DOCUMENT RESUME

ED 222 554	TM 820 708

AUTHOR	Mislevy, Robert J.; And Others
TITLE	Scale-Score Reporting of National Assessment Data (Final Report).
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Feb 82
GRANT	NIE-G-80-0003
NOTE	130p.; For related documents, see TM 820 707-712 and TM 820 716.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Educational Assessment; Item Analysis; *Item Sampling; *Latent Trait Theory; *Mathematics Achievement; National Surveys; Racial Differences; *Scaling; *Scores; Secondary Education; Sex Differences; Test Construction; Test Reliability
IDENTIFIERS	National Assessment of Educational Progress; *NIE ECS NAEP Item Development Project; Unidimensional Scaling; Unit of Analysis Problems

ABSTRACT

An approach was developed based on item-response models defined at the level of salient subject groups rather than at the level of individuals, designed for use with multiple-matrix sampling designs. In each of three National Assessment of Educational Progress (NAEP) mathematics subtopics, Reiser's group-effects latent trait model was fitted to the proportions of correct response to items as observed in the cells of a design. Item parameters and contrasts among demographic groups were estimated in each of four data sets: 1972-73 and 1977-78 data for 13-year-olds and 17-year-olds. Based on items common to two or more data sets, results were linked across ages and over time in each subtopic. Item parameters and group averages were obtained on scales common across ages and years. Successful calibration and linking in all subtopics demonstrates the feasibility of applying item-response methods to sparse sampling designs. However, scaling must be accomplished within fairly narrowly-defined skill areas, such as the NAEP subtopics, if the integrity of scales is to be maintained. Item response scaling of NAEP test booklets as a whole is discouraged. Primary type of information provided by the report: Results (Secondary Analysis). (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222554

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

SCALE SCORE REPORTING OF NATIONAL ASSESSMENT DATA

TM 820 708

SCALE-SCORE REPORTING OF NATIONAL ASSESSMENT DATA

(Final report)

Robert J. Mislevy
International Educational Services

Mark R. Reiser
Indiana University

Michele Zimowski
University of Chicago

February, 1982

The work upon which this publication is based is performed pursuant to Grant NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

SCALE-SCORE REPORTING OF NATIONAL ASSESSMENT DATA

Robert J. Mislevy, International Educational Services

Mark R. Reiser, Indiana University

Michele Zimowski, University of Chicago

Abstract

Perhaps the two most significant advances in educational measurement over the past twenty years have been item response theory and multiple-matrix sampling designs. Unfortunately, few researchers interested in assessment have been able to enjoy the full benefits of both advances simultaneously; the current methods of item response theory cannot deal with the sparse data (at the level of individuals) that characterize the most efficient sampling designs. This research develops an approach based on item-response models defined at the level of salient subject groups rather than at the level of individuals, designed for use with the most efficient multiple-matrix designs, i.e., those in which each sampled subject is presented at most one item per scale.

In each of three NAEP mathematics subtopics, Reiser's group-effects latent trait model was fit to the proportions of correct response to items as observed in the cells of a design including sex, race/ethnicity, region of the country, and size and type of community. Item parameters and contrasts among demographic groups were thus estimated in each of four age/year data sets: 1972/73 and 1977/78 data for 13-year olds and 17-year olds. (Data were taken from NAEP public release tapes from these age/years and from the NAEP mathematics 1972/78 "change" tape.) Based on items common to two or more age levels and/or assessment years, results were linked across ages and over time in each subtopic. Item parameters and group averages were then obtained on scales common across ages and years, despite the fact that different (but overlapping) sets of items had been administered in each age/year.

Successful calibration and linking in all three subtopics demonstrates the feasibility of applying item-response methods to the sparse sampling designs of modern assessment. It is seen, however, that scaling must be accomplished within fairly narrowly-defined skill areas, such as the NAEP subtopics, if the integrity of scales across demographic groups and over time is to be maintained. In particular, item response scaling of NAEP test booklets as a whole is to be most strongly discouraged as it virtually guarantees item parameter drift over time and poor fit to unidimensional item response models.

CONTENTS

PREFACE

I. INTRODUCTION AND BACKGROUND . . . . Page 1

The nature of educational assessment
Multiple-matrix sampling designs
Methods of reporting assessment results
Problem statement

II. ITEM RESPONSE-CURVE METHODS FOR ASSESSMENT. . . . . . Page 13

Introduction to item response-curve theory
Application to multiple-matrix samples of responses
Reiser's model for group effects
Linking results across assessments

III. EXAMPLES FROM NAEP'S ASSESSMENTS OF MATHEMATICS . . . Page 38

Introduction to the examples
Methodology
Results

IV. CONCLUSIONS . . . . Page 52

BIBLIOGRAPHY

APPENDICES

A. Technical presentation of Reiser's model for group effects
B. A minimum Chi-square solution for linking calibrations
   from forms with an arbitrary design of overlap
C. A computer program program for linking calibrations
D. Comments on the NAEP public-use data tapes

PREFACE

Item response-curve models have triggered no less than a revolution in educational measurement. Little wonder, since so many measurement problems that are difficult or impossible to solve within the framework of classical psychometric theory become quite tractable under the item response-curve approach; examples include analyses of the information that items and tests provide at various levels of ability, measurement on an invariant scale from any subset of of calibrated items, and simplified test-equating procedures.

To date these benefits have not been realized in the National Assessment of Educational Progress. The primary reason, perhaps, is NAEP's use of multiple-matrix sampling designs--efficient procedures guaranteed to provide economical estimates of group level attainment. Sufficiently precise estimates of group-level attainment may be obtained by administering only a few items from a given skill area to any selected subject. Unfortunately, the current state of item response-curve theory cannot handle data such as gathered by NAEP, wherein each subject responds to too few items in a specific skill area to permit the stable estimation of his ability level.

This project is intended to further the extension of item response-curve theory to the assessment setting. The foundations of the present work appeared in the estimation procedures outlined in Bock (1976) that were later put into practice with the California Assessment Program (Bock, 1979; Bock and Mislevy, 1981); item response curve models are in these applications defined not at the level of individuals but at the level of salient groups of individuals. Reiser's (1980) dissertation research introduced a group-level item response-curve model that is particularly suited to NAEP data, addressing characteristics of test items and performances in the cells of a design on persons. The present work develops procedures to link such results across assessment years and/or age groups. Examples are drawn from the 1972/73 and 1977/78 NAEP mathematics assessments.

CHAPTER I

INTRODUCTION AND BACKGROUND

The Nature of Educational Assessment

The purpose of educational assessment is to provide information about the levels of skills or attitudes in specified populations of subjects. Results may be compared from one population to another or from one point in time to another, in order to study the effects of educational treatments or societal trends. The distinguishing feature of assessments, however, is their focus on groups rather than on individuals.

By virtue of its distinct purposes, assessment requires a different technology than its close cousin, educational measurement. The "true-score" models of traditional psychometrics concern the measurement of individual subjects rather than groups of subjects; it is not surprising that the strategies of test construction designed to provide optimal measurement of individuals are not optimal for assessment. While borrowing heavily from the models and the concepts of educational measurement, assessment technology has gainfully employed ideas from other fields as well, notably those of opinion survey sampling and sampling design theory.

The state of the art of assessment in the United States is exemplified by the National Assessment of Educational Progress (NAEP) and the California Assessment Program (CAP). These two programs have, since their inceptions over a decade ago, been proving grounds for measurement and statistical advances designed to obtain efficient and economical estimates of group-level attainment. Our attention will focus primarily on the National Assessment, although discussion of certain topics will be clarified with examples from the California Assessment.

The National Assessment of Educational Progress charts levels of attainment in ten broad areas, including Reading, Science, Mathematics, and Writing Skills. Each area is assessed periodically, usually once every four or five years. Information is gathered mainly through the administration of multiple-choice and open-ended tasks from the target area to subjects selected in the NAEP sampling design. Demographic and educational background data are also obtained for each sampled subject. Results are reported as proportions of correct response to individual items and clusters of items, for groups of individuals defined by demographic variables such as age, sex, region of the country, size of community, and so on.

Comparing NAEP results over time or across age-groups requires measurement on an invariant scale. Proportions of correct response for a given item may be compared across all the assessments in which it was administered... but certainly trends in, say, Mathematics skill are inadequately revealed by performance on any single item. To overcome the idiosyncracies of individual items, information must be combined over several items testing the same essential skills.

Average percents-correct over clusters of items may instead be followed, but only as long as the composition of the cluster does not change. NAEP uses this option at present, but it is hampered by the fact that typically one fourth of the items in each assessment are released to the public and retired from the item pool. Comparisons across assessments of average percents-correct of clusters of items will become less reliable as the numbers of common items shrink over the years.

This report explores methods by which modern item response-curve measurement theory may be applied to the assessment setting to solve the problems of charting progress over time. The next section of this chapter reviews the basics of multiple-matrix sampling theory, the development which has contributed so much to the success of large-scale assessments such as NAEP to date. It is upon this sampling framework that measurement models must build if they are truly to advance the practice of assessment. Next, the current practices of reporting assessment results are reviewed. Their limitations and prospects for overcoming them are discussed. The chapter concludes with a succinct statement of the objectives of the present research.

Multiple-Matrix Sampling

The accountability movement of the 1960's inspired the creation of a number of local and statewide testing programs intended to provide feedback about the effects of public expenditures on education. The methodology employed in these programs was that of standardized achievement testing. Every pupil in a school or classroom was administered an achievement test consisting of as many as two hundred test items, an undertaking demanding hours or even days of classroom time from each pupil. Designed to provide maximal differentiation among students, these tests yield highly accurate scores for each pupil in but a few broad skill areas.

Averages of pupil-level scores obtained in such a scheme did indeed reflect levels of performance in the school or classroom, but in a most highly inefficient manner. The administration of intensive every-pupil testing with traditional achievement tests suffers several serious deficiencies if it is only the group-level results that are necessary for discussion by the public and the educational community. The large numbers of items which must be administered to a student in a skill area if distinctions are to be made among students are simply not necessary if only information about average levels of attainment in the group as a whole is desired. Such a scheme expends scarce educational resources to measure each student much more precisely than is required in an assessment, but by providing results in only a few broadly-conceived skill areas, offers little in the way of specific guidance for improving the curriculum.

During this same period, sample survey techniques were becoming a familiar and widely-accepted mechanism for guaging the strength of various attitudes and opinions among the public, mainly on issues of social or political relevance. Not every person is interviewed; not every person interviewed is asked all the same questions. Yet satisfactorily precise and reliable information is obtained about the prevalence of attitudes in the public at large. Why not apply these same methods to educational assessment?

At the request of William Turnbull, president of Educational Testing Services, Frederic Lord investigated the possibility of estimating levels of ability in a population by means of "multiple-matrix sampling"--that is, by administering different subsets of an item domain to different samples of persons (Lord, 1962; Lord and Novick, 1968, Chapter 11).

The simplest application of multiple-matrix sampling is in estimating the average item score in a population of N subjects for an item pool of K test items. The average score that would be obtained by administering every item to every subject can be approximated by observing the responses of, say, t different random samples of n subjects each to random samples of k items each. (This is referred to in the multiple-matrix sampling literature as a t/k/n design.) The expected value of the average item score over all such samples is the population average item score.

One of the most important results of Lord's investigations was the conclusion that the estimation of the population average is most precise for a given number of responses when k=1; that is, when the responses for different items have been obtained from non-overlapping samples of subjects. Stated simply, two responses contain more information about the population if they are from different persons than if they are from the same person.

Pandey and Carlson (1976), in a study of data from the California reading assessment, found the effect to be generalizable. The error variance associated with estimates of the population mean was reduced almost four-fold when, for the same number of responses, forms of ten items each were administered to samples of ten subjects each, as compared to a design under which the items were administered as two fifty-item forms to ten subjects each.

Practical work generally requires a more complicated sampling design than those described above. In the California Assessment, as an example, results must be reported individually for each school. In the National Assessment, results must be reported for the cells and the margins of a design based on demographic variables. Sampling of subjects must therefore be carried out within levels of stratification in both cases, allowing for the possibility of different selection probabilities within different cells to meet requirements for the precision of estimation.

Item pools are generally statified as well, into divisions of increasingly narrower skill requirements. The goal is to define classes of items which are similarly affected by specific attributes of educational treatments, in order that treatments can be monitored and modified as a result of the feedback from the assessment. Our attention in this paper will focus on items within the finest level of stratification of the item pool, which, following the practice of the California Assessment, we refer to as a "skill element."

Reporting Assessment Results

As noted above, comparisons of assessment results over time and across assessments requires measurement on an invariant scale. The method by which this requirement is achieved in public opinion survey research is to present subjects with questions that remain constant over time; by asking, for example, "Which of these candidates would you vote for if the election were today?" of subjects interviewed during the six months before the election, one may chart the flow of public support behind the candidates. Tyler's (1968) remarks and Womer's (1973) monograph suggest that this same method was originally intended for reporting the results of the National Assessment of Educational Progress.

The "fixed-item" approach to reporting the results of assessments, as it might be called, focuses on comparisons of performance between groups or across time on a single, specific task.

Interpretation is straight-forward as it applies to performance on that particular item, but the problem lies in generalizing the results. The 1972/73 NAEP Mathematics Assessment, for example, presented 200 items to 13-year-olds alone; results for the cells of a sex by race by size-and-type of community design would have to be expressed as some ten thousand separate percent-correct values. Comparisons across groups would vary across items as a result of measurement error as well as with the skills tapped by the items. How could such a preponderance of detail be suited to general public dissemination or discussion?

Educational test items, unlike public opinion survey questions, are not usually important in and of themselves, but instead as representatives of a class of tasks requiring similar skills. It is these generalized skills rather than the specific items that are addressed by instruction, and it is this level at which assessment results must be reported. The technology of educational measurement, as it had developed by the early 1970's (see, for example, Cronbach et al, 1972), was able to provide a framework for generalizing results across test items within a skill area: the "random-item" model.

Under the "random-item" approach to reporting the results of assessments, the specific items from a given skill area are considered a random sample from a population of items that, taken together, defines the area. The average item score by a group of subjects to a randomly-selected subset of these items is an estimate of the group's average for the entire population of items in the area. Because results are averaged over a number of items, peculiarities of item formats and distractors tend to cancel out, revealing trends which underly performance on all the items in the skill area. Under this model, average item scores may be compared across groups and over time even though different sets of items may have been administered, as long as the set of items administered in each assessment has been chosen at random from an invariant population of items.

The assumptions of the "random-item" approach are not, unfortunately, met in general practice. The problem lies with the requirement of randomly sampling items from an invariant item pool. If comparisons are desired across age groups, for example, subjects must be presented items from the same item pool; the efficiency of estimation suffers if younger subjects are presented just as many hard items as are necessary to tap the skills of older subjects, and older subjects are presented too many easy items just so the younger subjects can be tested.

A more serious problem is the charting of results over time. If item pools remain invariant, they cannot reflect new emphases in educational treatments nor can they retire items which have outlived their usefulness; neither can items be released to the public to aid the interpretation of results without compromising the integrity of the measurements. Yet if apparently desirable revisions to certain items are carried out, the average item scores estimated in different assessments are not comparable; they are estimates of performance in a shifting collection of items, perhaps harder or easier on the whole from one year to the next. Changes in subjects' skill levels are confounded with changes in the composition of the item pool.

The National Assessment, recognizing this problem, has responded by reporting results that are to be compared over time in two different ways.

For non-technical reports slated for public release, proportions of items correct are reported over all items in a skill area, despite modifications in the item pool and decidely non-random selections of items in an assessment. The comparisons implied by the figures for different years, although they do not meet the assumptions required to assure meaningfulness, are considered useful nonetheless. And indeed they may be good approximations of the comparisons that would have resulted under ideal conditions; i.e., true random sampling in both years from a fixed pool of items.

For scientific investigations, NAEP provides reports and data tapes based on only those items which appear on all the assessments to be compared. The assumptions of the "random-item" model may be met in this way, defining the skill area to be that which is measured by the average of that specific collection of items. Comparisons, restricted to these so-called "change items," suffer from the culling of items that cannot be matched, in that potentially useful information from these items must be ignored. The resources expended in gathering this information have not been justified in this respect. Analyses of trends over time are as well, as the set of items common to all assessments in question tends to shrink when more time points are considered.

Similar problems in the measurement of individual subjects have been overcome with the advent of item response-curve (IRC) measurement models. An IRC model individually parameterizes each test item in a suitable domain in terms of its relationship to an underlying scale of ability. Subjects may then be measured on an invariant scale of attainment, based on their responses to any subset--not just a randomly selected subset--of items. The challenge is to apply the methods of IRC theory to the setting of assessment, borrowing concepts and machinery to free reporting from the constraints of classical test theory, while at the same time building upon the multiple-matrix sampling framework.

Problem Statement

The objective of this research is to further the development of one approach to applying item response-curve methods to assessement data; namely, Reiser's (1980) group effects model, which (1) allows the estimation of group-level parameters from item responses obtained in an efficient multiple-matrix sampling design, (2) yields these parameter estimates on a scale that is invariant over time and across groups, and (3) permits the evolution of the item pool over time without degrading the integrity or the generality of the results. The steps we take to this end are as follows:

1. Develop an algorithm for linking estimates from the Reiser model across assessments. The approach will be based on a proposal by Tucker (1948). Linear transformations are determined to provide optimal agreement among estimates from an arbitrary number of assessments, linked by an arbitrary pattern of common items.

2. Demonstrate the use of the group-effects model and the linking program with data from the NAEP 1972/73 and 1977/78 Mathematics Assessments. Scales will be linked across assessment years and across the 13- and 17-year-old age groups in three skill areas.

CHAPTER II

ITEM RESPONSE-CURVE METHODS FOR ASSESSMENT DATA

This chapter develops an approach for adapting item response-curve methods to the assessment setting. The first section is a brief review of item response-curve theory as it has been developed for measuring individual subjects. Features and properties of IRC models that will be important in our generalization to group-level data will be emphacized. The second section discusses the notion of applying IRC methods to data obtained from multiple-matrix sampling designs. In particular we consider the option of defining IRC models at the level of subject groups rather than individuals. The third section is a non-technical description of Reiser's group-effects, an IRC model defined at the level of groups that is particularly well-suited to the demographic stratifications used by the National Assessment. The final section discusses the linking of results from the Reiser model from one assessment to others, thus providing the continuity of measurement necessary for longitudinal analyses. (The topics treated in the last two sections, the Reiser model and the linking procedures, are treated in a more technical manner in Appendices A and B respectively.)

Fundamentals of Item Response-Curve Theory

The models of item response-curve theory differ most radically from the models of traditional "true-score" psychometrics by parameterizing test items individually in terms of their relationships to the underlying ability, rather than treating them as random samples from a pool of interchangable items. Once a set of items has been "calibrated" (i.e., the parameters of the items have been estimated), a subject's ability can be estimated from his responses to any subset of the items. This is the case even when the items he has been presented are only easy ones or only hard ones--assuming that the IRC model fits the circumstances reasonably well.

The heart of an IRC model is a mathematical equation for the probability of a correct response to a particular item by a particular subject, in terms of one or more parameters that indicate the subject's abiliy and one or more parameters describing how responses to the item are influenced by ability.

To illustrate, we will consider the Birnbaum 2-parameter logistic item response-curve model, which will be seen to share many similarities with Reiser's group-effects model for assessment data. The probability that Subject i will respond correctly to Item j is given by the following function:

$$\mathrm{Prob}(X_{ij}=1) = \frac{\exp[1.7\ A_j\ (\theta_i - B_j)]}{1 + \exp[1.7\ A_j\ (\theta_i - B_j)]} \qquad (1)$$

where

X_{ij}, the response, is 1 if it is correct and 0 if not,

exp is the exponential function,

θ_i is the "ability" parameter of Subject i,

A_j is the "slope" parameter of Item j, and

B_j is the "threshold" parameter of Item j.

(The scaling constant 1.7 is included in this expression in order to make the item parameters in the Birnbaum logistic model match more closely the item parameters in the normal ogive model.)

The function shown above describes how likely it is that a subject with a given ability will respond correctly to Item j. This function can be graphed, as in Figure 1: the item response curve for Item j. It may be seen that subjects with very low values of θ have little chance of responding correctly. As θ increases, so do chances of responding correctly. For a subject with an ability that has the value Bj (the threshold of Item j), the chances of a correct response are 50-50. As ability continues to increase, chances of responding correctly increase also until it is nearly a certainly at very high levels of ability.

When this model is fitted to data, it is capable of accounting for the facts that--

(1) Some subjects perform better than others on the items in the skill area.

(2) Some items in the area are easier than others.

(3) Some items are more reliable indicators of the ability

than others.

Figure 2 shows three different item response curves on the same plot. It may be seen that, on the average, Item 3 is harder than Item 2, which is harder than Item 1. It may also be seen that the higher the value of an item's slope, the more sensitively an item reacts to changes in subject ability. Item 2 is more informative than Item 1, which in turn is more informative than Item 3.

The manner in which subject and item parameters combine to produce probabilities of correct response is illustrated in Tables 1 and 2. Table 1 shows, for the four hypothetical items and six hypothetical subjects, the quantity $1.7\ A_j\ (\theta_i - B_j)$. The orderly relationships among the parameter values are most clear in this chart, showing what are called the "logits" of correct response. Table 2 transforms these logits to the more familiar units of probabilities via Equation 1.

It may be noted at this point that the units for the subject and item parameters are unique only up to a linear transformation. That is, equivalent relationships may be expressed by transforming all the parameters by the linear function $f(x)=mx+b$ as follows:

$$\theta_i^* = m\ \theta_i + b$$

$$A_j^* = A_j\ /\ m$$

$$B_j^* = m\ B_j - b.$$

It may be readily verified that these transformed subject and item parameters yield exactly the same probabilities of correct response as the originals, since

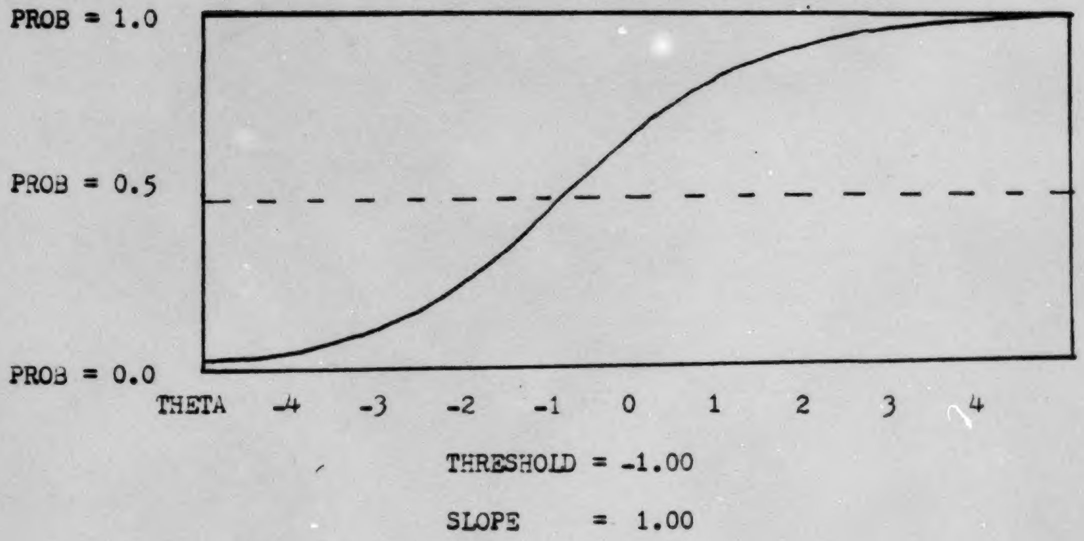

FIGURE 1
AN ITEM CHARACTERISTIC CURVE

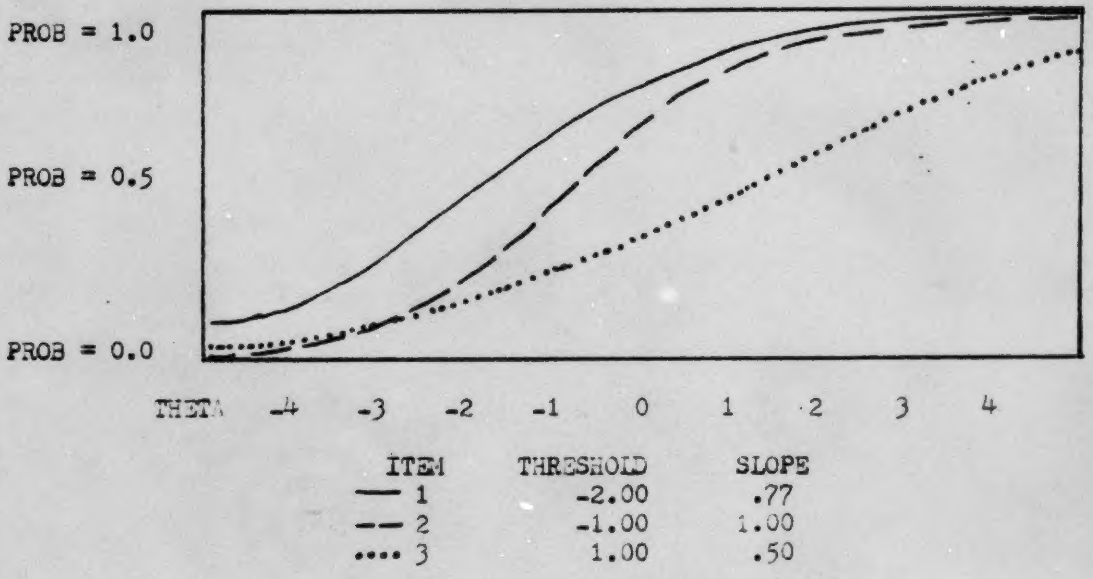

FIGURE 2
SEVERAL ITEM CHARACTERISTIC CURVES

TABLE 1
LOGITS OF EXPECTED PROPORTIONS CORRECT

			ITEMS			
		Bj:	-2.000	-1.000	1.000	2.000
SUBJECT	θi	1.7 Aj:	1.000	0.500	0.500	1.000
1	-0.500		1.500	0.250	-0.750	-2.500
2	-1.500		0.500	-0.250	-1.250	-3.500
3	0.500		2.500	0.750	-0.250	-1.500
4	-0.500		1.500	0.250	-0.750	-2.500
5	1.000		3.000	1.000	0.000	-1.000
6	0.000		2.000	0.500	0.500	-2.000

TABLE 2
EXPECTED PROPORTIONS CORRECT

			ITEMS			
		Bj:	-2.000	-1.000	1.000	2.000
SUBJECT	θi	1.7 Aj:	1.000	0.500	0.500	1.000
1	-0.500		.818	.562	.321	.076
2	-1.500		.622	.438	.223	.029
3	0.500		.924	.679	.438	.182
4	-0.500		.818	.562	.321	.076
5	1.000		.953	.731	.500	.269
6	0.000		.881	.622	.378	.119

$$1.7\, A_j^* (\theta_i^* - B_j^*) = 1.7\, A_j (\theta_i - B_j).$$

The implication is that when parameters are estimated for the same set of items from different sets of data, they can be expected to differ by such as linear transformation. It is a practical problem to estimate the optimal transformation; various solutions and proposals have been made by Tucker (1948), Lord and Novick (1968), and Haebera (1981).

Benefits of Item Response Curve Models

When a collection of subjects' responses to test items can be adequately summarized in terms of an item response-curve model like the one described above, several benefits accrue:

Invariance with respect to item selection. Once a collection of items has been calibrated (i.e., the parameters of the items have been estimated), a subject's ability may be estimated on the basis of his responses to any subset of the items--randomly selected or not. This means that, as an example, younger students may be administered mostly easy items while older students are administered mostly difficult items from the same scale.

New items can be added to the domain. New items measuring the same trait can be linked into an existing scale by administering them along with items that have already been calibrated. The new items can be calibrated from this data, then their fit to the model verified before they are used to estimate subjects' abilities.

Flawed items can be corrected. Items found to have flaws in their grammar, format, or conception can be revised, then re-calibrated into the domain as if they were new items.

Items can be dropped from the bank. Without affecting the the scale of measurement, items may be retired from use, either because they are outdated or because they will be used to illustrate the content area in reports released to the public.

Content-referencing of scores. The scaled scores (θ) from IRC theory are defined implicitly by the probabilities of correct responses they imply for each of the items in the skill area. The meaning of an ability estimate can be interpreted, therefore, by inspecting the content of the items with thresholds in that region of the scale--without reference to the distribution of ability in any population of subjects. Scale scores may still be interpreted in the more familiar manner of norm-referencing of course, with the compuation of percentiles, stanines, standard scores, and so on, with regard to specified populations of subjects.

Linearity with external variables. Because IRC ability estimates are not subject to the floor and ceiling effects of numbers-right and percents-correct, they tend to have more linear relationships with external variables such as SES, age, and years of education.

Well-defined standard errors. Because items are large-sample calibrated, they are considered 'fixed' rather than 'random' from a statistical point of view. For this reason, standard errors of estimating subjects' abilities, and, indirectly, reliabilities, are easy to compute (see Lord & Novick, 1968; Mislevy, 1981). Moreover, these standard errors are correctly expressed as a function of the ability itself rather than gratuitously and erroneously assumed constant as in classical theory for number-right scores.

Suitability for longitudinal studies. With the use of alternate test forms consisting of items from the same scale, IRC ability estimates are amenable to the study of trends or program effects.

Assumptions of Item Response Curve Models

If an IRC model is to fit data, the assumptions of the model must be reasonably well satisfied. The main assumptions are discussed below.

Unidimensionality. Nearly all applied work uses IRC models that assume a single underlying ability scale. This means that subjects' differing probabilities of correct response, with respect to each of the items in the scale, can be described by a single variable. If one subject's probability of a correct response to a given item is higher than that of a second subject, the assumption of unidimensionality implies that the first subject has higher probabilities of correct response than the second subject on all the items in the scale.

Local (conditional) independence. A subject's response to a given item is assumed to depend only on his level of ability, not on extraneous factors such as the position of the item on the test or his responses and reactions to preceding items.

Temporal stability. Item parameters, and equivalently, relationships among items, must remain stable over time to guarantee the comparability of ability estimates over time.

Goodness-of-fit. The item and person parameters of the IRC model must accurately account for the probability of a correct response to any item from any subject who is to be measured. This is equivalent to saying that the item parameters and the scale of ability they imply must be invariant over subjects. (Experience has shown that the more homogeneous the content of the items, the more likely it is that this assumption will be satisfied.)

Satisfying these assumptions is more of a skill than a science. During the past decade, practitioners have begun to build up the body of experience necessary to apply item response curve theory at the level of measuring individuals. Still questions remain, concerning topics such as the range of ability over which item parameters can retain the same values and the possibility that item parameters may 'drift' over time.

With the exception of the California Assessment Program, there has been little experience to date with the problem of meeting the assumptions of IRC models in the context of the sparse (at the level of individuals) samples of item responses gathered in efficient multiple-matrix designs. Guidelines derived from their experience will be discussed in the following section, and employed in the examples in the following chapter.

Application to Multiple-Matrix Samples of Responses

Clearly, the advantages of item response curve theory offer considerable benefit to educational assessment. Not only can the restriction of a fixed item bank from which items must be drawn at random be lifted, but results can be reported on a content-referenced scale: estimates of levels of attainment can be interpreted in terms of probabilities of correct response to the items whose thresholds define a scale.

The main obstacle to the application of IRC theory to the assessment setting is, ironically, the efficient design of the multiple-matrix samples. IRC theory, as presently conceived, has been designed for the measurement of individual subjects. To estimate the ability of an individual subject with IRC methods, several items from the scale must be administered to him. This practice is at odds with the aim of multiple-matrix sampling, which provides economical information about groups by eliminating the measurement of individual subjects.

There are three approaches by which the technology of IRC theory can be applied to multiple-matrix samples of responses. The following paragraphs consider each in turn.

Subject-level model, subject-level estimates. The first approach by which IRC methods may be applied to multiple-matrix samples of item responses employs an IRC model like those descibed in the previous section, modelling the probabilities of correct response of individual subjects. Each subject sampled for a given skill area is administered enough items from that skill area to permit the estimation of his ability. The resulting estimates of the abilities of individual subjects may then be averaged over subpopulations as desired.

Several benefits of IRC theory may be enjoyed under this approach. First, the necessary computational methods are available, having been developed over the past decade for use in measuring individuals. Second, the restrictions on the item bank are relieved; items could be dropped from the bank or new ones could be calibrated in. Third, content-referencing of score estimates is possible. And fourth, random selection of items for test forms is no longer required; harder and easier forms could be developed and administered appropriately, so that the items an individual takes may be more informative about him and, consequently, about the subpopulations to which he belongs.

As noted above, however, this approach will require the administration of several items from that area--perhaps as many as fifteen or twenty--proscribing the use of the most efficient multiple-matrix designs. (Pandey and Carlson [1976] demonstrate a 380-percent increase in efficiency for estimating a group average using ten-item forms as compared to fifty-item forms, observing the same number of responses in both designs.) If the application of IRC theory is truly to advance the state of the art of assessment, it must build upon the advances already gained through the use of efficient sampling designs rather than discard them.

Subject-level model, group-level estimates. A second approach is to define an IRC model at the level of individual subjects, but to estimate the parameters of the distributions of ability in subpopulations directly, without estimating the abilities of individual subjects.

This approach is well conceived for application to the most efficient multiple-matrix designs. Each sampled subject in the assessment must be administered only one or two items in any skill area. It is possible to estimate the parameters of the distributions of ability in any subpopulation on the basis of the responses of the subjects sampled from that subpopulation, and to estimate relationships among skill areas or between skills and external measured variables--all without estimating the ability of any individual subject.

Efficient methods of estimation under this approach are still under development. The rudiments for one method are found in Andersen and Madsen (1977) and Sanathanan and Blumenthal (1978), which discuss the estimation of population parameters from subject-level data under the restrictions that all subjects have been administered the same set of items and the IRC model is the one-parameter logistic. Extension to the general IRC case and to multiple-matrix data are given by Mislevy (1982).

Group-level model, group-level estimates. The third approach, the one upon which the present research is based, defines an item response curve model at the level of subject groups rather than at the level of individual subjects. One or more item parameters still relate each item in a skill area to an underlying scale of attainment, but the ability (or attainment) parameters are for groups of subjects rather than individuals. Rather than modelling the probabilities of correct responses from specific individuals, a group-level IRC expresses the probability of a correct response to a particular item from a subject selected at random from groups at the various levels of attainment. A group ability parameter may thus be interpreted as the average over the subjects in that group.

A group-level IRC would be defined at the lowest level of stratification of the population of subjects for which results are to be studied or reported. In the California Assessment, for example, an IRC is defined at the level of schools; school-level score estimates are then averaged to the levels of districts, Los Angeles areas, counties, and the state as a whole when desired.

The sampling scheme upon which such an IRC model is based is the most efficient multiple-matrix sampling design, in which each sampled subject responds to at most one item from any given skill area. Under this design and in a group-level model, the responses of the individual subjects from a given group may be considered independent, given the ability parameter of the group. In this way the IRC assumption of local independence is satisfied.

Group-level IRC's can be justified in two different ways. First, they may be seen simply as models for data analysis which may be used profitably when they are able, with their item and group parameters, to describe the matrix of item-by-group proportions of correct response in the skill area under consideration. In this sense they are a generalization of logistic models for the analysis of binary data (Cox, 1970; Bock, 1975), with any interaction terms between "item" factors and "subject-group" factors are constrained to follow the patterns describable as item response curves with possibly different slope parameters. Second, group-level IRC's may be seen as an integration over group distributions of phenomena described by subject-level IRC's. Under this interpretation, the distributions in all groups are assumed identical in shape, and may differ only as to location.

In two special cases a simple relationship exists between item parameters from the group-level IRC and those from a subject-level IRC. (1) The distributions of ability within groups may be considered to be concentrated on a single point (Bock, 1976), in which case the item parameters in the group-level IRC would be identical to those in the subject-level IRC. This case assumes that the grouping of subjects accounts for all systematic variation among them. (2) The distributions of abilities may be assumed normal within groups, and, if a normal ogive subject-level IRC is assumed, the group-level item threshold and slope parameters are functions of the subject-level item parameters and the common dispersion of ability within groups (Mislevy, 1982).

Defining Skill Areas for Assessment

Comparing attainment over time or across subpopulations requires item parameters that remain stable over time and across subject groups. This requirement is most easily achieved in the setting of individual measurement by scaling within skill areas defined narrowly rather than broadly.

This prescription can be a burden in the setting of individual measurement, since it implies that each individual to be measured must be administered several items from each of several skill areas, defined narrowly to guarantee stable item parameters but highly correlated in the population of individuals.

The same prescription can be a boon in the setting of assessment. In efficient approaches of IRC application to assessment, each subject will be administered only a few items from each separately scaled skill area. This means that the number of skill areas which can be measured at the level of groups can be very large, without requiring excessive time for administration. The Grade 3 California Assessment, for example, measures school-level attainment in 61 skill elements, while requiring less than an hour from each subject. (The items from these skill elements are distributed among thirty different test forms, each of which contains thirty-four items from different elements; each sampled subject is administered one randomly selected test form.)

The manner in which skill elements are selected for separate scaling is based on the requirement for stable relationships among items' relative difficulties. The California Assessment Program has attempted to define skill elements in terms of educational practice: if skill elements are based on "indivisible curricular elements", all items in a scale will be similarly affected by curricular change. Changes in a school's performance over time, then, may appear as increases in one element and decreases in another, but will be consistent with respect to all the items within an element. Because progress (or lack of progress) can be monitored at the level at which educational treatments are applied, CAP results help school officials adjust the balance of emphases of various components of the curriculum.

Reiser's Model for Group Effects

The National Assessment of Educational Progress (NAEP) employs sampling at both the item and the subject level. The assessment instrument in a given content area is constructed of items sampling specified objectives and assigned to one of a number of forms, with the number of forms varying across age levels and content areas in accordance with the number of objectives in which performance is to be measured. Each such form is administered to a national probability sample of approximately 2500 persons, selected by the cluster method from the appropriate age group (9-, 13-, or 17-year olds). Pupils, designated by age rather than grade, are tested by NAEP personnel outside the classroom. All pupils in any one testing session are administered the same form, and are "paced" through it by a tape recording that determines the amount of time spent on each item. Free-response as well as multiple-choice items are used.

The results of these tests are aggregated not to the level of schools, as in the California Assessment, but to the cells of a multi-way demographic classification of subjects. Ways of classification include age, sex, racial/ethnic group, size and type of community, region of the country, and parental education. The emphasis is on measuring progress (change) in attainment of the objectives as seen in the population as a whole and in the subpopulations defined by the demographic classifications. Typically each assessment deals with one or more content areas, each of which is typically assessed every four or five years.

At present NAEP does not make use of any type of scale-score reporting. Results are expressed as percents-correct for items slated for public release, or as average percents-correct over the items in an objective or a content area. Because of the difficulties mentioned above with the interpretation of these averages when the item pool changes from year to year, NAEP could make use of the item-invariant scales offered by IRC models.

In dissertation research, Reiser (1980) generalized a model by Bock (1976) to provide a group-level IRC appropriate to the aims and the current practices of NAEP assessments. The Reiser model is based on the following assumptions:

1. Each objective is scaled separately. The items within an objective are considered sufficiently homogeneous to function as what is referred to in the California Assessment as a "indivisible curricular unit." That is, differences between subpopulations and changes over time may differ across objectives, but are essentially the same for all the items measuring a given objective.

2. Each item representing an objective will appear on a different test form. (In present NAEP assessments, this assumption is not strictly satisfied; occasionally two or three items from the same objective appear on the same form.)

3. The distributions of ability within each of the cells of the classification scheme have the same shape, differing at most by location (i.e., cell average levels of attainment). The demographic classification is assumed to absorb all variation between schools or other levels of clustering in the sampling design.

4. The expected proportion of correct responses to an item within the ultimate subclasses of the demographic classes is a two-parameter logistic function of the parameters of the item and demographic effects for that cell.

An Example

Perhaps the best way to introduce Reiser's model is with a (relatively) simple numerical example. We begin by saying that the form of the model is very similar to that of the 2-parameter logistic IRC model described above, except that the focus is not explaining the probability of a correct response from a specified subject, but for a subject selected at random from a specified group (i.e., a cell from the demographic classification scheme).

Consider four items from a skill objective and the six cells of a sex-by-age design, including ages 9, 13, and 17. Assume that pilot-testing of the items has indicated their relative difficulties. To each age group, only the two items at the appropriate level of have been administered: Items 1 and 2 to 9-year olds, Items 2 and 3 to 13-year olds, and items 3 and 4 to 17-year olds. For each item targetted for a given age level, random samples of subjects from each sex are administered the item. All sex-by-age-by-item samples are equal in size. Suppose that the proportions of correct response observed in this administration are as shown in Table 3.

Now the comparison of average percents-correct over all the items taken suggests a decline in attainment as age increases, from .610 to .500 to .317. This result is clearly an artifact of the design of administration. It is obvious in this example that average percents-correct cannot be compared across sets of items that differ in difficulty.

One alternative is to compare age groups on the basis of the items that they have taken in common. Item 2, for example, shows .500 correct for 9-year olds and .621 correct for 13-year olds; Item 3 shows .380 correct for 13-year olds and .439 for 17-year olds. These comparisons, illustrating increasing levels of performance with increasing age, are valid but inefficient; each is based on only half the data available from the age groups being compared. Moreover, no such comparison can be made between 9- and 17-year olds, because they have taken no items in common.

The first step in understanding Reiser's model is to consider the logits of these proportions, as shown in Table 4. The model attempts to explain these values as functions of item parameters Bj (threshold) and Aj (slope), and cell average attainment (Θkl, where k designates sex and l designates age). The form of the model is as follows:

$$Ljkl = 1.7\ Aj\ (\Theta kl - Bj),$$

where $Ljkl$ represents the logit of the proportion of correct responses to Item j from the cell with sex designation k and age

TABLE 3
OBSERVED PROPORTIONS CORRECT

		ITEMS				
AGE	SEX	1	2	3	4	"AVERAGE"
9	F	.818	.562	-	-	} .610
9	M	.622	.438	-	-	
13	F	-	.679	.438	-	} .500
13	M	-	.562	.321	-	
17	F	-	-	.500	.269	} .317
17	M	-	-	.378	.119	

TABLE 4
LOGITS OF EXPECTED PROPORTIONS CORRECT

				ITEMS			
			Bj:	-2.000	-1.000	1.000	2.000
AGE	SEX	θkl	1.7 Aj:	1.000	0.500	0.500	1.000
9	F	-0.500		1.500	0.250	(-0.750)	(-2.500)
9	M	-1.500		0.500	-0.250	(-1.250)	(-3.500)
13	F	0.500		(2.500)	0.750	-0.250	(-1.500)
13	M	-0.500		(1.500)	0.250	-0.750	(-2.500)
17	F	1.000		(3.000)	(1.000)	0.000	-1.000
17	M	0.000		(2.000)	(0.500)	0.500	-2.000

designation 1. In terms of proportions correct, the logits are transformed as follows:

$$Pjkl = \frac{\exp(Ljkl)}{1 + \exp(Ljkl)} \tag{2}$$

From the observed logits of correct response, item and group parameters must be estimated. In Reiser's model, as with all 2-parameter logistic IRC models, there are two linear dependencies that must be resolved arbitrarily--it is this fact that permits all parameters to be rescaled by a linear transformation as discussed above. In this example we resolve them by restricting the average of the thresholds of the items to be one and the distance between the highest and lowest thresholds to be four. Under these constraints, the estimates of the item and group parameters are as follows:

Item	Threshold	Slope
1	-2.00	1.00
2	-1.00	0.50
3	1.00	0.50
4	2.00	1.00

Age	Sex	Ability
9	F	-0.50
9	M	-1.50
13	F	0.50
13	M	-0.50
17	F	1.00
17	M	0.00

As befits an artificial example, these estimates perfectly account for the observed proportions of correct response as shown in Table 3, when combined via Equation 2. An examination of the ability, or scale score, values for the demographic groups shows a clear increase in levels of attainment with increasing age, from -1.00 to 0.00 to 0.50, with males and females averaged in each age group. Moreover, this pattern accounts for the differences between ages for all items. The comparison is thus based on all the observations.

The second step in understanding the Reiser model requires a closer look at the scale scores of the six sex-by-age cells. As noted above, score averages for age groups with sexes combined are -1.00, 0.00, and 0.50. Score averages for sex groups with age groups combined are 0.50 for females and -0.50 for males. Together these age and sex marginal effects account for each of individual cells; that is, there is no sex-by-age interaction. To obtain the scale score of any cell, three steps are required:

1. Start with an initial approximation of 0.00.
2. To account for the age effect, subtract 1.00 if the cell is for 9-year olds and add 0.50 if it is for 17-year olds.
3. To account for the sex effect, subtract 0.50 if the cell is for males and add 0.50 if it is for females.

A distinguishing feature of Reiser's model is that the levels of ability in the ultimate subgroups in the design need not be estimated individually, but may be expressed as functions of some smaller number of effects related to the ways of classification.

Statistical tests for the presence of effects, both main and interaction, are easily obtained by comparing how well various nested models explain the observed proportions of correct item responses across the cells of the design.

Reiser's dissertation research, as an example, used a classification scheme based on sex, race, and size-and-type of community (STOC). The analysis concerned Skill in Computing Fractions, with data from the 1977/78 assessment of 13-year olds. He found that the variation among the attainment levels of the cells in this 2-by-3-by-7 design could be explained in terms of just main effects for the three variables and race-by-sex interaction.

The parameters of Reiser's model may be estimated by the method of maximum likelihood: An equation like Equation 2 above expresses the probability of a correct response to a given item from a given cell in the design. The product of these expressions over all the items and cells, appropriately weighted to reflect the numbers of attempts each observed proportion represents, is the probability of the entire data set, as a function of item and group-effect parameters. Item and group-effect parameters are then found that maximize this probability. (See Appendix A for a more technical description of the model and the estimation procedures.)

Linking Results Across Assessments

The Reiser model outlined above has the capacity for analyzing multiple-matrix samples of item responses with the item-invariance properties that distinguish the IRC approach. Previous use of the model (Reiser's dissertation) considered data from one time point only, considering just the proportions of correct response to all items in an objective as observed in the cells of a demographic classification of subjects. But charting results over time is the raison d'etre of assessment; capabilities for linking the results of assessments from different points in time is essential to any method of analyzing such data.

In principle it is possible to analyze simultaneously data from several points in time with the Reiser model. All that is necessary is the (possibly incomplete) matrix of proportions of correct response to the items in the objective in question, from each cell in the demographic classification of subjects, at each point in time. The analysis proceeds as described in the previous section, except that the effects which constitute constraints in modelled cell probabilities now include a main effect for time and, if desired, interactions of time and demographic effects (i.e., allowing for the measurement of differential progress in different subpopulations).

This approach has in fact been carried out in the present study, with data for two points in time within a single age group. The geometric increase in the number of item-by-group cells as additional time points are considered, however, leads to an exponential increase in the computing resources necessary to estimate the parameters in the model. Clearly this approach is not well suited to longitudinal analyses of any complexity.

A more managable approach is to estimate the item and group-effect parameters from each point in time separately, then link the results on the basis of items that are common across time points. If the assumptions of the model are correct, the item parameters for the linking items in two assessments should differ by only a linear transformation:

$$Aj^* = Aj / m$$

$$Bj^* = m\ Bj - b,$$

where the linear transformation $f(x)=mx+b$ translates the item parameters from the second point in time to the base scale. The same transformation is then applied to the ability estimates of the subject groups and group effects. It is necessary, then, to be able to estimate values of m and b which will make the each item's response lines from the two time points match most closely after the item slopes and thresholds from the second time point are appropriately transformed.

Methods of estimating m and b have been proposed by Tucker (1948), Lord and Novick (1968), and Haebara (1981). One simple approach is to calculate the mean and the standard deviation of the item thresholds at both points in time, then choose m and b so that the mean and standard deviation of the rescaled Time II thresholds matches the corresponding values from Time I. That is,

$$m = S(I) / S(II)$$

$$b = [S(I)/S(II)]\ \overline{X}(II) + \overline{X}(I),$$

where S(k) denotes the standard deviation of the thresholds at Time k and $\overline{X}(k)$ denotes their mean.

This simple procedure does not take into account the fact that some item parameters may be estimated more accurately than others, either because more subjects have responded to a particular item at a particular point in time or because the item is more closely matched to the average of the ability in the population of subjects. Moreover, linking is based on information from threshold estimates only, ingnoring potentially useful information from item slope estimates.

A more sophisticated linking procedure which takes both of these factors into account is described in Appendix B. The procedure is designed to link any number of calibrations, as long as the data from all calibrations are linked by patterns of common items. It is not necessary for any item to appear on all calibrations, but each calibration must share at least two items with other calibrations, and each calibration must be at least indirectly linked with all other calibrations. (Calibration a is directly linked with Calibration b if they have an item in common. Calibration a is indirectly linked with Calibration z if there is a sequence of directly linked calibrations beginning with a and ending with z.)

CHAPTER III

EXAMPLES FROM THE NAEP MATHEMATICS ASSESSMENTS, 1972/73 AND 1977/78

Introduction to the Examples

The process of constructing scales affording the implementation of the aforementioned methods began with a perusal of the NAEP classification scheme of unreleased items. Three skill element categories comprised of sufficient numbers of items, common to all cells yet appearing in unique booklets within each cell of the age/year breakdown, were located. The NAEP classifications satisfying the criteria were Understanding Mathematical Concepts, i.e., value 4 of Cognitive Subtopics, Arithmetic Computation, and Algebraic Manipulations, i.e., values 1 and 5, respectively, of Mathematical Skills Subtopics. Tables 5, 6, and 7 present the NAEP identification numbers of the items in these scales, along with their locations in the various age/year assessment forms.

While items in the first category require the ability to translate from one form of symbolism or language to another, those in the other demand the rote application of the learned methods of arithmetic and algebra. Hence, the examples illustrate the application of the methods to measures of rudimentary as well as abstract levels of mathematical ability.

TABLE 5
DISTRIBUTION OF ITEMS:
UNDERSTANDING MATHEMATICAL CONCEPTS

	13-YEAR OLDS				17-YEAR OLDS			
	1977/78		1972/73		1977/78		1972/73	
NAEP #	FORM	ITEM	FORM	ITEM	FORM	ITEM	FORM	ITEM
5-A45532					1	S0109		
5-B41532					2	S0204		
5-B41732					3	S0315		
5-B31732	8	T0823			4	S0426		
5-N00002	1	T0141	1	T0124	5	S0503	5	S0509
5-B11008	6	T0607	6	T0633	6	S0639	6	S0621
5-A71043	7	T0712			7	S0718		
5-A21022	2	T0206	2	T0221	8	S0830	8	S0806
5-B32632	10	T1020			9	S0921		
5-K30004	4	T0431	4	T0402	10	S1039	(NOT USED)	
5-K10010	9	T0908	9	T0926	11	S1106	11	S1125
5-B33232	3	T0319						
5-G43009	5	T0540						
5-H12025			7	T0733			7	S0707
5-G20001							10	S1001
5-K51020			5	T0502				
5-B22011			8	T0816				
5-A21032							4	S0423

TABLE 6
DISTRIBUTION OF ITEMS: ALGEBRAIC MANIPULATIONS

	13-YEAR OLDS				17-YEAR OLDS			
	1977/78		1972/73		1977/78		1972/73	
NAEP #	FORM	ITEM	FORM	ITEM	FORM	ITEM	FORM	ITEM
5-H11025	2	T0202	2	T0208	1	S0139	1	S0125
5-G10003	3	T0337	3	T0305	9	S0906	9	S0916
5-H11007	6	T0603	6	T0619	2	S0202	2	S0204
5-C50022	8	T0807	8	T0818	4	S0402	4	S0402
5-G43005					3	S0338	3	S0323
5-H11015					7	S0707	7	S0718
5-G44007					8	S0829	8	S0804
5-I31001					11	S1102	11	S1104
5-B21325					5	S0538		
5-B20925					10	S1031		
5-B20125					6	S0605		
5-H11002	4	T0432	4	T0410				
5-B40225	7	T0706A						
5-B30425	9	T0939A						
5-H11010			5	T0524			5	S0504
5-H11026			1	T0121				
5-H21001							6	S0606

TABLE 7
DISTRIBUTION OF ITEMS:
ARITHMETIC COMPUTATION

	13-YEAR OLDS				17-YEAR OLDS			
	1977/78		1972/73		1977/78		1972/73	
NAEP #	FORM	ITEM	FORM	ITEM	FORM	ITEM	FORM	ITEM
5-B13002					1	S0135		
5-C30010	8	T0810	8	T0828	2	S0205	2	S0212
5-C10049	7	T0733	7	T0701	4	S0406	4	S0421
5-C20006	2	T0203	2	T0210	5	S0504	5	S0512
5-A23009	4	T0435	4	T0424	6	S0631	6	S0602
5-F30006					7	S0708	7	S0727
5-F00006	5	T0537	5	T0516	8	S0835	8	S0825
5-A31732					10	S1023		
5-B31225					11	S1132		
5-A11832	1	T0125						
5-A45232	3	T0327A						
5-A22010	6	T0602	6	T0605				
5-C10009	9	T0902	9	T0901				
5-A34632	10	T1027						
5-C10011			1	T0126				
5-F00007							1	S0104
5-C20021			3	T0320			11	S1118
5-F00003							10	S1008
5-C20022							9	S0902
5-C30001							3	S0306

Within the 1977/78 assessment year completion of the scales, that is, selection of one item per remaining booklet, was accomplished through reference to the three NAEP classifications. To maximize the number of possible between cell comparisons, items common to other cells of the age/year breakdown were granted priority in the selection process.

Because the item classification schema of the 1972/73 assessment differed from that of the 1977/78 asessment, the selection of items was based on an item-by-item scrutiny of the available pool. Once again, items common to other cells were given selection priority.

The resulting scales vary in the total number of items as well as in the number of among-cell item communalities. For example, Understanding Mathematical Concepts is defined by a total of 17 items of similar content. The number of items within any one cell of the age/year breakdown ranges between 7 and 11; pairs of cells share between 3 and 6 items. Likewise, the Arithmetic Computation scale is comprised of a total of 20 items, the number of items within any cell falling in the interval of 9 to 11, the between-cell communalities ranging from 5 to 7 items. Finally, a total of 17 items define the Algebraic Manipulation scale, the number of items within each cell varying from 7 to 11, the number of shared items varying from 4 to 8.

Within each cell subject groups are defined according to a multi-way demographic classification. The cross-classification is based on four variables, namely sex, race, size-and-type of community, and region of the country.

Methodology

In order to obtain item parameter and subgroup effect estimates on a common scale across years and age groups, the following steps were taken in each of the three skill areas:

1. Fit the Reiser group-effects model to data from each age/year separately.
2. Establish unit-size and location of scale with respect to the results of 1977 13-year olds.
3. Determine optimal linear transformations of remaining age/year results to reference scale.
4. Transform item parameter and group-effect estimates to reference scale.

The remainder of this section amplifies these procedures.

Step 1: Fit Group-Effects Model to each Age/Year Separately

The basic data addressed by the Reiser group-effects model are the counts of numbers of attempts and numbers of correct responses to each item observed in each cell of the design on persons. The classification of persons used in these examples is based on sex (male and female), race (Hispanic, Black, and white), region of the country (Northeast, Southeast, Central, and West), and STOC, or size and type of community (extreme rural, low metropolitan, small places, main big cities, urban fringe, medium cities, and high metropolitan). The design consists of 168 cells in all. Data from persons with missing data in any of these variables or not identified in one of the three main racial/ethnic categories was excluded from the analyses.

Numbers of attempts and correct responses to each item in a skill area were accumulated for each cell in the design, with each person's data weighted in proportion to his NAEP sampling weight. Weights were rescaled so that the sum of weights was equal to the number of observations; in this way oversampling was taken into account but numbers of observations were not exaggerated.

In its attempt to explain the observed (weighted) proportions of correct response to each item from each cell in the design on persons, Reiser's model yields estimates for threshold and slope parameters for each item (reflecting items' relative difficulties and reliabilities) and for contrasts among selected cells in the design on persons. A maximum of 168 contrasts could be estimated with the present design, including all main effects and all possible interactions. Because Reiser's dissertation research suggested that interactions were generally negligible, only main effects were included here. Simple contrasts were employed for sex, race, and region:

Male - Female

Hispanic - white

Black - white

Northeast - West

Southeast - West

Central - West

So-called identity contrasts were employed for STOC. Conditional on the effects listed above, the average scale-score in each STOC category is estimated. The location and unit-size, which must be arbitrarily specified, were provisionally set by fixing the "extereme rural" effect at -1.00 and the "high metro" effect at +1.00.

The parameter estimates obtained in a given run of the Reiser model, then, consist of thresholds and slopes for the items presented in that age/year, one sex effect, two race effects, three region effects, and seven STOC effects. Each estimate is accompanied by a large-sample standard error of estimation, except for the two STOC effects that were fixed to set the scale.

Item parameter and subject-group effects can be combined to produce estimated proportions of correct response to each item in each cell. Tests of fit are obtained by comparing these estimated proportions with the observed proportions. Likelihood ratio Chi squares have been provided for each run, with numbers of degrees of freedom equal to the numbers of non-empty cells times the numbers of items presented in the age/year in question, minus the number of parameters estimated in the run. Because likelihood ratio Chi squares can be questionable for small cells--and some cells in the design, such as high metro female Hispanics in the Northeast, are very small--the more robust Freeman-Tukey Chi squares are also provided for selected runs for comparison.

Step 2: Establish Reference Scale in 1977 13-Year Old Results

The size of units and the zero point of the scale must be arbitrarily fixed in the Reiser group-effects model. The scale for these examples has been set by requiring the estimated grand mean of 1977 13-year old results to be zero and the distance between the "extreme rural" and "high metro" STOC categories to be two.

As noted above, the provisional scales for each age/year run were set by requiring the values for these two STOC categories to be -1.00 and +1.00 respectively, so the unit-size in the 1977 13-year old provisional scale meets specification. The grand mean over all 1977 13-year olds was determined by averaging STOC effects, each weighted by the proportion of the population it represented. This grand mean was subtracted from all 1977 13-year old STOC effects and item thresholds so as to fix the grand mean at zero. This scaling is the reference to which the remaining age/year results will be transformed.

Step 3: Determine Linear Transformations for Remaining Age/Years

Under the assumption that the items in a scale define the same variable across ages and over years, the sets of item threshold estimates for items presented in two age/years will differ by only a linear transformation, aside for random errors of estimation. Similarly, the two sets of item slope estimates will differ non-randomly by a scaling constant only, namely the scaling constant required in the linear transformation of the item threshold estimates. Once the linear transformation has been determined, item parameters and group effects may be put onto a common scale.

The weighted least-squares algorithm described in Appendix B has been used to obtain optimal estimates of the linear transformations required to bring the results from the remaining age/years to the reference scale established for the 1977 13-year olds. Information is utilized from all occurances of an item in two or more age/years, including the precision with which each estimate is determined. The goal of the algorithm may be described as minimizing the squared weighted differences among item parameters estimated in two or more age/years.

It has been determined that the 1977 13-year old results are the reference scale, so the identity transformation is known to be appropriate for that age/year. Estimation error variation of the rescaling constants has been apportioned across all four age/years, however, to reflect uncertainty in all age years in the transformation of group-effect estimates. Table 8 displays the estimates and standard errors of estimation used in the examples.

TABLE 8
RESCALING PARAMETERS

AGE/YEAR	SLOPE	SE	INTERCEPT	SE
UNDERSTANDING MATHEMATICAL CONCEPTS				
1977 13-YEAR OLDS	1.000	.035	.473	.118
1972 13-YEAR OLDS	.686	.026	.885	.099
1977 17-YEAR OLDS	1.187	.045	2.801	.127
1972 17-YEAR OLDS	.744	.031	2.957	.125
ALGEBRAIC MANIPULATION				
1977 13-YEAR OLDS	1.000	.022	.432	.067
1972 13-YEAR OLDS	.329	.014	.609	.071
1977 17-YEAR OLDS	.824	.021	2.254	.066
1972 17-YEAR OLDS	.849	.021	2.612	.065
ARITHMETIC COMPUTATION				
1977 13-YEAR OLDS	1.000	.021	.614	.078
1972 13-YEAR OLDS	.762	.021	.100	.078
1977 17-YEAR OLDS	.577	.018	2.888	.068
1972 17-YEAR OLDS	.675	.022	3.443	.087

Step 4: Transform Results to Reference Scale

Let f(x)=mx+b be the estimated linear transformation of the results for a given age/year to the reference scale. The transformation of STOC effects and the grand average, reflecting locations along the scale, are accomplished as follows:

$$\theta^* = m\,\theta + b$$

$$SE(\theta^*) = \mathrm{Sqrt}[(m^2\ SE^2(\theta) + \theta^2\ SE^2(m) + SE^2(b)]$$

(The adjustment of the standard error neglects a term attributable to the covariance of the errors of estimation of m and b, as these terms have been found to be negligible.) The transformations of sex, race, and region effects, which represent distances along the scale, are accomplished by:

$$\theta^* = m\,\theta$$

$$SE(\theta^*) = \mathrm{Sqrt}[(m^2\ SE^2(\theta) + \theta^2\ SE^2(m)]$$

Final estimates of item parameters were obtained by first transforming the threshold estimates in each age/year in the same manner as STOC effects and slope estimates in the same manner as contrast effects, and then obtaining weighted threshold and slope averages for each item over all ages and years in which it was administered.

Taken together, the final estimates of item parameters and group effects can be used to compute expected proportions of correct response to any item in the scale from any cell in the design on persons. To facilitate the interpretation of the effects, additional tables of conditional margins have been provided; that is, estimated averages for each of the levels in the sex, race, and region factors, under the assumption of "all other factors held constant." The average of the conditional effects over all the levels of a given factor in a given age/year, with each level weighted in accordance with the proportion of the population it represents, is the grand mean for that year. The marginal proportions of the factors used in these computations are given in Table 9.

Results

The results of the procedures outlined above are summarized in Tables 10 through 18 and Figures 3 through 5. Tables 10 through 12 and Figure 3 concern Understanding Mathematical Concepts: Table 10 presents rescaled item parameter estimates from all four age/years and grand averages, Table 11 presents the corresponding estimates of group effects, Table 12 presents the conditional margins they imply, and Figure 3 plots item thresholds and race/ethnicity averages against the ability scale. Similar information for Algebraic Manipulations is presented in Tables 13 through 15 and Figure 4, and for Arithmetic Computation in Tables 16 through 18 and Figure 5. Highlights are discussed below.

Overall indices of goodness-of-fit of the group-effects model to data from each age/year for Concepts, Manipulation, and Computation are found in Tables 11, 14, and 17 respectively. Chi-

TABLE 9
SAMPLED MARGINAL PROPORTIONS

SUBGROUP	1977, AGE 13	1972, AGE 13	1977, AGE 17	1972, AGE 17
MALE	.499	.505	.487	.521
FEMALE	.501	.496	.513	.479
HISPANIC	.060	.056	.046	.043
BLACK	.164	.167	.138	.152
WHITE	.776	.777	.816	.805
NORTHEAST	.227	.248	.232	.244
SOUTHEAST	.226	.256	.229	.254
CENTRAL	.316	.248	.327	.253
WEST	.231	.247	.213	.249
EXTREME RURAL	.099	.101	.100	.101
LOW METRO	.101	.101	.099	.098
SMALL PLACES	.332	.333	.349	.364
URBAN FRINGE	.154	.106	.157	.084
MAIN BIG CITY	.141	.120	.136	.115
MEDIUM CITY	.071	.140	.058	.139
HIGH METRO	.102	.099	.101	.099

NOTES: 1. DATA FROM APPROXIMATELY 24,000 PERSONS IS ANALYZED IN EACH AGE/YEAR.

2. PROPORTIONS SHOWN ABOVE INCORPORATE NAEP CASE WEIGHTS.

TABLE 10

ITEM PARAMETER ESTIMATES: MATHEMATICS CONCEPTS

	1977, AGE 13				1972, AGE 13				1977, AGE 17				1972, AGE 17				GRAND AVERAGES			
ITEM	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE
5-A45532									1.01	0.24	0.29	0.03					1.01	0.24	0.29	0.03
5-B41532									2.31	0.25	0.14	0.02					2.31	0.25	0.14	0.02
5-B41732										***										
5-B31732	3.05	0.36	0.19	0.01					2.22	0.23	0.16	0.02					2.46	0.19	0.19	0.01
5-N00002	-1.69	0.36	0.20	0.03	-1.49	0.42	0.29	0.04	-1.26	0.57	0.20	0.03		*			-1.54	0.25	0.24	0.02
5-B11008	0.18	0.22	0.16	0.02	-1.18	0.49	0.15	0.03	-0.21	0.43	0.20	0.03	-0.97	0.72	0.19	0.03	-0.13	0.18	0.18	0.01
5-A71043		*							1.17	0.27	0.21	0.03					1.17	0.27	0.21	0.03
5-A21022	0.41	0.16	0.25	0.03					1.24	0.23	0.28	0.03	0.98	0.34	0.26	0.04	0.72	0.12	0.26	0.02
5-B32632	2.68	0.32	0.18	0.03						*							2.68	0.32	0.18	0.03
5-K30004	1.11	0.27	0.10	0.02	1.44	0.20	0.19	0.03	0.04	0.42	0.18	0.02					1.16	0.15	0.17	0.02
5-K10010	3.76	0.44	0.19	0.03	3.35	0.42	0.26	0.04					3.55	0.19	0.21	0.03	3.54	0.16	0.22	0.02
5-B33232	0.23	0.22	0.15	0.02													0.23	0.22	0.15	0.02
5-G43009	1.92	0.28	0.14	0.02													1.92	0.28	0.14	0.02
5-H12025					1.37	0.20	0.19	0.03					1.41	0.30	0.21	0.03	1.38	0.17	0.20	0.02
5-G20001													4.62	0.25	0.25	0.04	4.62	0.25	0.25	0.04
5-K51020					0.78	0.14	0.31	0.04									0.78	0.14	0.31	0.04
5-A21032													4.10	0.24	0.20	0.03	4.10	0.24	0.20	0.03
5-B22011						**														

* ITEM DELETED; QUESTIONABLE DATA ON NAEP PUBLIC RELEASE TAPE.

** ITEM DELETED; CONVERGENCE PROBLEMS IN GROUP-EFFECTS PROGRAM.

*** ITEM DELETED; ESTIMATED THRESHOLD VALUE TOO EXTREME.

TABLE 11
ESTIMATES OF GROUP EFFECTS: UNDERSTANDING MATHEMATICAL CONCEPTS

EFFECT	AGE 13, 1977		AGE 13, 1972		AGE 17, 1977		AGE 17, 1972	
GRAND MEAN	.00	(.14)	.36	(.12)	1.87	(.15)	2.30	(.15)
MALE-FEMALE	-.07	(.09)	.71	(.14)	.32	(.10)	.56	(.12)
HISP-WHITE	-2.39	(.35)	-1.76	(.34)	-2.20	(.32)	-2.36	(.42)
BLACK-WHITE	-2.93	(.37)	-2.55	(.42)	-2.92	(.34)	-2.98	(.46)
NE-WEST	.66	(.16)	.00	(.13)	.67	(.17)	.17	(.13)
SE-WEST	-.16	(.16)	-.27	(.14)	-.13	(.15)	-.18	(.14)
CENTRAL-WEST	.67	(.14)	.26	(.14)	.72	(.15)	.34	(.13)
EXTREME RURAL	-.53	(.12)	.20	(.10)	1.61	(.14)	2.21	(.13)
LOW METRO	-.71	(.27)	-.14	(.26)	.83	(.30)	1.57	(.29)
SMALL PLACES	-.25	(.19)	.54	(.16)	1.71	(.21)	2.16	(.20)
MAIN BIG CITY	.03	(.21)	.52	(.20)	2.07	(.22)	2.50	(.23)
URBAN FRINGE	.00	(.21)	.67	(.19)	2.22	(.22)	2.42	(.22)
MEDIUM CITY	.69	(.23)	.50	(.19)	2.73	(.27)	2.58	(.20)
HIGH METRO	1.47	(.13)	1.57	(.10)	3.99	(.14)	3.70	(.13)
CHI SQUARE (LR)	1697.69		1367.79		1628.06		1518.32	
CHI SQUARE (FT)	NA		NA		1478.31		1082.20	
DEGREES FREEDOM	1178.00		954.00		1147.00		808.00	

TABLE 12
ESTIMATED CONDITIONAL MARGINS: UNDERSTANDING MATHEMATICAL CONCEPTS

SUBGROUP	AGE 13, 1977	AGE 13, 1972	AGE 17, 1977	AGE 17, 1972
GRAND MEAN	.00	.55	2.03	2.38
MALE	-.04	.91	2.19	2.66
FEMALE	.04	.19	1.87	2.10
HISPANIC	-1.77	-.68	.33	.57
BLACK	-2.31	-1.47	-.38	-.05
WHITE	.62	1.07	2.53	2.94
NORTHEAST	.33	.55	2.34	2.47
SOUTHEAST	-.49	.28	1.54	2.12
CENTRAL	.34	.81	2.39	2.64
WEST	-.33	.55	1.67	2.30
EXTREME RURAL	-.53	.20	1.61	2.21
LOW METRO	-.71	-.14	.83	1.57
SMALL PLACES	-.25	.54	1.71	2.16
MAIN BIG CITY	.03	.52	2.07	2.50
URBAN FRINGE	.00	.67	2.22	2.42
MEDIUM CITY	.69	.50	2.73	2.58
HIGH METRO	1.47	1.57	3.99	3.70

```
AGE-RACE/ETHNICITY   θ     ITEM                ABBREVIATED TEXT
----------------------------------------------------------------------------------------------
                    6.0
                     |
                     |
                     |
                     |
                     |    5-G20001   $Y / (4 BOYS) = ?
                     |
                    4.0   5-A21032   IF N IS ODD, N+1 IS EVEN
                     |
                     |    5-K10010   SEGMENT XY= 1/2 x 4 INCHES, OR 2 INCHES
                     |
17-W                 |
                     |    5-B32632   IF A*B=(AxB)-B, 4*5=(4x5)-5 OR 15
          17-W       |
                     |    5-B31732   ANY NUMBER TIMES ONE IS THAT NUMBER
                    2.0   5-G43009   TEMPLATE FOR ASSOCIATIVE PRINCIPLE HOLDS FOR BOTH + AND X  *
                     |    5-H12025   IF X<4, X+7<11
                     |    5-A45532   NEGATIVE NUMBER DIVIDED BY POSITIVE NUMBER IS NEGATIVE
13-W                 |    5-K30004   LINE SEGMENT HM TWICE AS LONG AS NP
                     |    5-A21022   EVEN NUMBER + 2 IS EVEN  *
17-H      13-W       |    5-K51020   DISTANCE BETWEEN CENTERS
          17-H       |    5-B33232   IF Z<6 AND Y<Z THEN Y<6  *
17-B                0.0   5-B11008   A>5 & B>5 INSUFFICIENT INFO. FOR RELATION OF A AND B
          17-B       |
13-H                 |
                     |
                     |
13-B                 |    5-N00002   IF HENRY>BILL AND BILL>PETE, THEN HENRY>PETE
          13-H       |
                   -2.0
          13-B       |
----------------------------------------------------------------------------------------------
1972      1977
```

* ITEM TEXT SLIGHTLY REVISED IN ORDER TO MAINTAIN SECURITY.

FIGURE 3

ITEM THRESHOLDS AND RACE/ETHNICITY CONDITIONAL MARGINS:
UNDERSTANDING MATHEMATICAL CONCEPTS

TABLE 13

ITEM PARAMETER ESTIMATES: ALGEBRAIC MANIPULATIONS

	1977, AGE 13				1972, AGE 13				1977, AGE 17				1972, AGE 17				GRAND AVERAGES			
ITEM	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE
5-H11025	0.08	0.15	0.25	0.03					0.01	0.31	0.31	0.04	0.30	0.32	0.30	0.04	0.10	0.12	0.29	0.02
5-G10003	5.46	0.67	0.29	0.04	4.72	1.00	0.32	0.08		*			2.92	0.16	0.20	0.03	3.10	0.16	0.26	0.02
5-H11007	0.53	0.12	0.34	0.04	0.52	0.11	0.42	0.10	0.90	0.24	0.29	0.04	0.86	0.23	0.36	0.04	0.59	0.07	0.34	0.02
5-G50022	6.12	0.99	0.15	0.03	3.93	0.81	0.26	0.07	2.98	0.15	0.23	0.03	3.36	0.15	0.25	0.03	3.21	0.11	0.23	0.02
5-G43005									5.22	0.35	0.32	0.04	4.84	0.25	0.34	0.04	4.96	0.20	0.33	0.03
5-H11015									5.69	0.39	0.25	0.03	4.81	0.28	0.24	0.03	5.11	0.23	0.24	0.02
5-G44007									6.18	0.49	0.27	0.04	6.62	0.50	0.25	0.04	6.39	0.35	0.26	0.03
5-I31001									6.63	0.56	0.72	0.13	8.89	1.26	0.42	0.10	7.00	0.51	0.65	0.09
5-H11010					2.11	0.36	0.31	0.07					0.68	0.28	0.25	0.03	1.23	0.22	0.27	0.03
5-H11002	2.17	0.19	0.35	0.04	2.96	0.56	0.29	0.07									2.25	0.18	0.34	0.03
5-H11026					-2.16	0.69	0.32	0.08									-2.16	0.69	0.32	0.08
5-B40225	-0.77	0.20	0.30	0.03													-0.77	0.20	0.30	0.03
5-B30425	3.99	0.45	0.22	0.03													3.99	0.45	0.22	0.03
5-H21001													5.75	0.36	0.31	0.04	5.75	0.36	0.31	0.04
5-B20925									8.24	0.96	0.37	0.07					8.24	0.96	0.37	0.07
5-B21325										**										
5-B20125										**										

* ITEM DELETED: QUESTIONABLE DATA ON NAEP PUBLIC RELEASE TAPE.

** ITEM DELETED: CONVERGENCE PROBLEMS IN GROUP-EFFECTS PROGRAM.

TABLE 14
ESTIMATES OF GROUP EFFECTS: ALGEBRAIC MANIPULATION

EFFECT	AGE 13, 1977		AGE 13, 1972		AGE 17, 1977		AGE 17, 1972	
GRAND MEAN	.00	(.09)	.36	(.08)	1.87	(.08)	2.30	(.08)
MALE-FEMALE	-.26	(.08)	-.35	(.11)	.25	(.08)	.25	(.07)
HISP-WHITE	-1.81	(.24)	-1.29	(.33)	-2.18	(.32)	-1.85	(.26)
BLACK-WHITE	-1.84	(.19)	-1.88	(.44)	-2.41	(.29)	-2.57	(.27)
NE-WEST	.33	(.12)	1.13	(.29)	.88	(.16)	.65	(.12)
SE-WEST	-.34	(.13)	.29	(.12)	-.20	(.13)	-.08	(.10)
CENTRAL-WEST	.25	(.11)	.81	(.21)	.33	(.09)	.02	(.10)
EXTREME RURAL	-.57	(.07)	.28	(.07)	1.43	(.07)	1.77	(.07)
LOW METRO	-.91	(.25)	-.58	(.32)	.77	(.27)	1.75	(.17)
SMALL PLACES	-.29	(.14)	-.13	(.12)	1.70	(.14)	2.21	(.12)
MAIN BIG CITY	-.08	(.16)	-.22	(.16)	2.12	(.15)	2.38	(.15)
URBAN FRINGE	.67	(.15)	-.30	(.16)	1.95	(.15)	2.20	(.15)
MEDIUM CITY	.24	(.17)	.47	(.14)	2.60	(.18)	2.49	(.13)
HIGH METRO	1.43	(.07)	.94	(.07)	3.08	(.07)	3.47	(.07)
CHI SQUARE (LR)	4504.31		2146.00		2857.26		1527.28	
CHI SQUARE (FT)	1257.52		1275.63		980.67		1349.41	
DEGREES FREEDOM	814.00		814.00		921.00		1161.00	

TABLE 15
ESTIMATED CONDITIONAL MARGINS: ALGEBRAIC MANIPULATION

SUBGROUP	AGE 13, 1977	AGE 13, 1972	AGE 17, 1977	AGE 17, 1972
GRAND MEAN	.00	.36	1.87	2.29
MALE	-.13	.19	2.00	2.42
FEMALE	.13	.53	1.75	2.16
HISPANIC	-1.40	-.54	.12	.91
BLACK	-1.43	-1.14	-.11	.19
WHITE	.41	.75	2.30	2.77
NORTHEAST	.25	.94	2.49	2.80
SOUTHEAST	-.42	.10	1.41	2.07
CENTRAL	.17	.61	1.93	2.17
WEST	-.08	-.20	1.61	2.15
EXTREME RURAL	-.57	.28	1.43	1.77
LOW METRO	-.90	-.58	.77	1.75
SMALL PLACES	-.29	-.13	1.70	2.21
MAIN BIG CITY	-.08	-.22	2.12	2.38
URBAN FRINGE	.67	-.30	1.95	2.20
MEDIUM CITY	.24	.47	2.60	2.49
HIGH METRO	1.43	.94	3.08	3.47

```
AGE-RACE/ETHNICITY   θ     ITEM                 ABBREVIATED TEXT
------------------------------------------------------------------------------------
                    8.0  5-B20925   IF N=3K AND N+K=72, THEN K=18 AND N=54
                     |
                     |
                     |   5-I31001   POINTS (X,Y) ON CIRCLE SATISFY X SQ + Y SQ = 36
                     |
                     |   5-G44007   FACTORS OF X SQUARE - 5X + 6 ARE (X-2) AND (X-3)
                    6.0
                     |   5-H21001   FIND SOLUTION SET OF (X-1)(X+7)=0
                     |
                     |   5-H11015   IF 3X + 6 - 14 = X + 2 THEN X=5
                     |   5-G43005   (2X-1)(X+3)= 2X SQUARE + 5X - 3  *
                     |
                     |
                    4.0  5-B30425   3X + 5Y + 4X = 7X + 5Y  *
                     |
                     |
                     |   5-C50022   IF A/B = C/D, THEN AxD = BxC IS TRUE
                     |   5-G10003   1/3 x A/2 = A/6
17-W                 |
                     |
       17-W          |   5-H11002   5 IN BOX MAKES 3(BOX +6)=21 TRUE   *
                    2.0
                     |
                     |   5-H11010   IF 3X-3 = 12 THEN X=?
                     |
17-H                 |
13-W                 |
       13-W          |   5-H11007   IF 2/3 = X/15 THEN X = 10   *
17-B   17-H          |   5-H11025   IF X+2 > 7, X MUST BE > 5   *
       17-B         0.0
                     |
13-H                 |
                     |   5-B40225   THE VALUE OF X+6 WHEN X=3 IS 9   *
13-B   13-H          |
       13-B          |
                     |
                   -2.0
                     |   5-H11026   IF X-3 = 7, THEN X=?
------------------------------------------------------------------------------------
1972   1977
```

* ITEM TEXT SLIGHTLY REVISED IN ORDER TO MAINTAIN SECURITY.

FIGURE 4

ITEM THRESHOLDS AND RACE/ETHNICITY CONDITIONAL MARGINS: ALGEBRAIC MANIPULATIONS

TABLE 16

ITEM PARAMETER ESTIMATES: ARITHMETIC COMPUTATION

	1977, AGE 13				1972, AGE 13				1977, AGE 17				1972, AGE 17				GRAND AVERAGES			
ITEM	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE	THRESH	SE	SLOPE	SE
5-C20006	1.40	0.16	0.20	0.02	2.04	0.26	0.28	0.04	1.73	0.20	0.38	0.05	1.11	0.35	0.30	0.04	1.58	0.11	0.29	0.02
5-F00006	3.12	0.29	0.24	0.03	2.74	0.35	0.22	0.03	2.81	0.14	0.26	0.04		*			2.86	0.12	0.24	0.02
5-C10049	-0.86	0.23	0.25	0.03	-0.51	0.18	0.24	0.03	-0.85	0.68	0.24	0.04	-0.38	0.61	0.26	0.04	-0.63	0.13	0.25	0.02
5-C30010	1.28	0.15	0.22	0.03	1.59	0.22	0.22	0.03	0.69	0.42	0.21	0.04	1.31	0.35	0.26	0.04	1.32	0.11	0.23	0.02
5-A23009	3.41	0.34	0.20	0.03	2.09	0.28	0.22	0.03	3.14	0.16	0.20	0.03	3.22	0.16	0.23	0.03	3.06	0.10	0.21	0.02
5-A22010	5.42	0.58	0.26	0.04	4.79	0.60	0.25	0.04									5.11	0.42	0.26	0.03
5-C10009	-4.75	0.90	0.17	0.03	-6.63	1.29	0.12	0.02									-5.37	0.74	0.15	0.02
5-F30006									6.14	0.45	0.30	0.04		*			6.14	0.45	0.30	0.04
5-A45232	3.49	0.34	0.20	0.03													3.49	0.34	0.20	0.03
5-A34632	5.60	0.70	0.16	0.02													5.60	0.70	0.16	0.02
5-A11832	-1.60	0.28	0.29	0.03													-1.60	0.28	0.29	0.03
5-B31225									3.95	0.18	0.32	0.05					3.95	0.18	0.32	0.05
5-B13002									0.71	0.33	0.40	0.06					0.71	0.33	0.40	0.06
5-A31732									4.21	0.22	0.26	0.04					4.21	0.22	0.26	0.04
5-F00007													6.35	0.41	0.27	0.04	6.35	0.41	0.27	0.04
5-C20021					2.08	0.26	0.26	0.03					1.37	0.37	0.20	0.03	1.84	0.21	0.24	0.02
5-F00003													1.08	1.09	0.19	0.03	1.08	1.09	0.19	0.03
5-C20022													-0.70	0.91	0.13	0.03	-0.70	0.91	0.13	0.03
5-C10011					-5.37	0.94	0.15	0.02									-5.37	0.94	0.15	0.02
5-C30001														**						

* ITEM DELETED; QUESTIONABLE DATA ON NAEP PUBLIC RELEASE TAPE.
** ITEM DELETED; CONVERGENCE PROBLEMS IN GROUP-EFFECTS PROGRAM.

TABLE 17
ESTIMATES OF GROUP EFFECTS: ARITHMETIC COMPUTATION

EFFECT	AGE 13, 1977		AGE 13, 1972		AGE 17, 1977		AGE 17, 1972	
GRAND MEAN	.00	(.17)	-.19	(.16)	2.43	(.12)	3.02	(.15)
MALE-FEMALE	-.13	(.08)	-.20	(.08)	.33	(.08)	.35	(.08)
HISP-WHITE	-2.51	(.30)	-2.17	(.32)	-1.68	(.27)	-1.69	(.29)
BLACK-WHITE	-2.61	(.25)	-3.38	(.42)	-2.15	(.29)	-2.62	(.35)
NE-WEST	.33	(.13)	1.13	(.19)	.68	(.14)	.62	(.13)
SE-WEST	-.92	(.16)	.03	(.12)	-.05	(.12)	.20	(.10)
CENTRAL-WEST	.00	(.12)	.82	(.15)	.34	(.10)	.50	(.11)
EXTREME RURAL	-.39	(.08)	-.66	(.08)	2.31	(.07)	2.77	(.09)
LOW METRO	-.60	(.25)	-1.47	(.27)	1.65	(.23)	2.31	(.22)
SMALL PLACES	-.25	(.16)	-.36	(.15)	2.20	(.14)	2.93	(.14)
MAIN BIG CITY	-.44	(.19)	-.09	(.17)	2.54	(.14)	3.10	(.17)
URBAN FRINGE	.58	(.16)	.35	(.16)	2.60	(.14)	3.08	(.16)
MEDIUM CITY	.02	(.19)	.19	(.16)	2.89	(.16)	3.10	(.15)
HIGH METRO	1.61	(.08)	.86	(.08)	3.47	(.07)	4.12	(.09)
CHI SQUARE (LR)		NA		2082.84		3105.16		1569.53
CHI SQUARE (FT)		NA		1584.01		1268.41		1417.59
DEGREES FREEDOM		NA		1064.00		911.00		1150.00

TABLE 18
ESTIMATED CONDITIONAL MARGINS:
ARITHMETIC COMPUTATION

SUBGROUP	AGE 13, 1977	AGE 13, 1972	AGE 17, 1977	AGE 17, 1972
GRAND MEAN	.00	-.19	2.43	3.02
MALE	-.07	.29	2.60	3.19
FEMALE	.07	-.09	2.27	2.84
HISPANIC	-1.93	-1.67	1.13	1.80
BLACK	-2.03	-2.88	.66	.88
WHITE	.58	.49	2.80	3.49
NORTHEAST	.46	.45	2.85	3.02
SOUTHEAST	-.79	-.65	2.13	2.90
CENTRAL	.14	.14	2.52	3.19
WEST	.13	-.68	2.18	2.69
EXTREME RURAL	-.39	-.66	2.31	2.77
LOW METRO	-.60	-1.47	1.65	2.31
SMALL PLACES	-.25	-.37	2.20	2.93
MAIN BIG CITY	-.44	-.09	2.54	3.10
URBAN FRINGE	.58	.35	2.60	3.08
MEDIUM CITY	.02	.19	2.89	3.10
HIGH METRO	1.61	.86	3.47	4.12

```
AGE-RACE/ETHNICITY   θ     ITEM              ABBREVIATED [illegible]
---------------------------------------------------------------------------------
                     |
                     |   5-F00007   3**0 = ?
                    6.0  5-F30006   6 IS THE WHOLE NUMBER NEAREST SQRT OF 38  *
                     |
                     |   5-A34632   3 20/15 IS NEXT STEP OF SUBTRACTION PROBLEM
                     |   5-A22010   (3)(-3) + 4 = -5
                     |
                     |
                    4.0  5-A31732   LEAST COMMON DENOMINATOR OF 7/15 & 4/9 IS 45
                     |   5-B31225   2 TIMES SQRT 5 = SQRT 20  *
17-W                 |   5-A45232   300.00/36 IS FIRST STEP FOR 3 DIVIDED BY .36
                     |
                     |   5-A23009   EXPRESS 9/100 AS 9%
            17-W     |
                     |
                     |
17-H                2.0  5-C20021   1/2 + 1/3 = ?
                     |   5-C20006   2/3 OF 9 = 6
                     |   5-C30010   (.4) x (3.6) = 1.44  *
            17-H     |   5-F00003   4**3 = ?
17-B        13-W     |   5-B13002   (+3) + (-3) = 0  *
13-W        17-B     |
                     |
                    0.0
                     |
                     |   5-C10049   420 DIVIDED BY 35 = 12  *
                     |   5-C20022   (1/2)(1/4) = ?
                     |
                     |
13-H                 |   5-A11832   3/9 IS THE SAME AS 1/3  *
            13-H   -2.0
            13-B     |
                     |
13-B                 |
                     :
                     |
                   -5.0  5-C10011   SUM OF FOUR NUMBERS
                     |   5-C10009   43 + 71 + 75 + 92 = 281
---------------------------------------------------------------------------------
1972        1977
```

* ITEM TEXT SLIGHTLY REVISED IN ORDER TO MAINTAIN SECURITY.

FIGURE 5

ITEM THRESHOLDS AND RACE/ETHNICITY CONDITIONAL MARGINS: ARITHMETIC COMPUTATION squares less than twice their degrees of freedom are considered indicative of acceptable fit; it may be seen, however, that several of the likelihood ratio (LR) Chi-squares exceed this value. Freeman-Tukey (FT) Chi-squares, on the other hand, range between one and one-and-a-half times their degrees of freedom, suggesting a highly satisfactory goodness-of-fit. Inasmuch as the two indices are asymptotically equivalent but the Freeman-Tukey Chi-square is less susceptible to problems with small cells, it would appear that the observed proportions of correct response in the examples are well-explained by the group effects model and the parameter estimates.

It will be recalled that an item's threshold is the point along the ability scale at which we would expect 50-percent correct responses to the item. Group averages may be interpreted in terms of item content, then, by inspecting the content of the items in the region of the scale at which the average falls. The group's proportion of correct responses would be about 50-percent for items in that neighborhood, less than 50-percent for items with higher thresholds, and greater than 50-percent for items with lower thresholds. In this way the content of items with thresholds at various points along the scale forms a picture of the ability scale upon which group effects are measured.

Figures 3 through 5, depicting the example scales, show reasonable patterns of increasing complex or advanced item content at increasing levels of θ. Algebraic Manipulations and Arithmetic

Computation show a broader and more evenly-spaced distribution of items along the scales than does Understanding Mathematical Concepts. Items from the latter scale are more concentrated in the area that includes average 13-year olds and 17-year olds, but more sparse in the lower regions of the scale.

Under the assumptions of the model, item parameters in a scale are invariant across ages and assessment years. If this is true, progress may be charted in terms of ability estimates alone; changes in the value of the global ability correctly reflect changes in probabilities of correct response to each individual item in the scale. Departures from this assumption, such as varying change over years from one item to another, are revealed as discrepencies among an item's parameter estimates in different age/years, after optimal rescaling (Tables 10, 13, and 16).

An examination of these tables shows few age/year item parameters further than one-and-a-half standard errors of estimation from the corresponding grand averages; in other words, the assumption of invariant item parameters acorss the ages and years in the examples is reasonably well satisfied. The interpretation of cases in which certain items were unexpectedly hard or easy in a particular age/year are left to curricular experts, although one pattern is suggested in the results for 1977 13-year olds in Algebraic Manipulations: both items found unexpectedly difficult in this age/year, compared to results on the other items in the scale, deal with solving fractions equations. In the main, however, the assumption of invariant scales across age/years and the subsequent discussion of trends in terms of ability estimates rather than for individual items are justifyable.

The universal test score decline of the seventies spans the period covered by our examples, and with minor exceptions, appears in all skill areas, age levels, and demographic subgroups addressed here. Only in the area of Arithmetic Computation and only for 13-year olds did levels of performance increase. In Concepts and Manipulation, equal decline was observed at both ages.

Male versus female contrasts in all three skill areas exhibit an interesting age-by-sex interaction: 13-year old females outperform 13-year old males, but 17-year old males outperform 17-year old females. (An exception is 1972 Concepts, where 13-year old males outperform females). One possible explanation of this result is that the well-established superiority of males in certain areas of mathematics (Anastasi, 1958) is manifest in the more abstract tasks in the higher regions of the scales but overwhelmed by superior study habits of females in the elementary grades on the less abstract tasks in the lower regions of the scales.

Race/ethnicity contrasts uniformly exhibit highest levels of performance by whites, followed at a distance by Hispanics then Blacks. The magnitude of the difference is such that the averages of 13-year old whites equal or exceed those of 17-year old blacks. A comparison of 1972 and 1977 results shows blacks at both age levels catching up somewhat in Arithmetic Compuation but both black and Hispanic 13-year olds falling further behind in Understanding Mathematical Concepts. In the remaining ages and skill areas, relative positions among the race/ethnicity groups remained about the same.

Contrasts among different regions of the country are of a much smaller magnitude. Performance is highest in the Central region, generally followed by the Northeast, West, and Southeast. The period covered by the examples saw a shift of population from the Northeast and Central regions to the Southeast and West; possible correlates of this shift are visible in region contrasts and margins. In Concepts, the distance between the Northeast and Central averages and the Southeast and West averages increased at both age levels from 1972 to 1977. Similar gaps in Manipulation decreased for 13-year olds but increased for 17-year olds; gaps in Compuation also decreased for 13-year olds but remained unchanged for 17-year olds.

The results for size and type of community (STOC) show the effects of a high concentration of well-educated and highly-paid professionals on the level of achievement in a neighborhood. The low metropolitan areas have a low level of income and few professional reside in them; hence, the level of achievement is low. Levels of income and proportions of professionals rise as one goes form low metropolitan areas to rural areas, small places, main big cities, and to urban fringe areas. Finally, in urban areas where the levels of income and education are highest, young peoples' levels of performance are highest also.

Declines in performance were generally more pronounced in the STOC categories that were lowest to begin with--i.e., low metropolitan and rural areas--but less pronounced in the higher STOC categories. In fact, the high metropolitan category showed increases as often as declines, particularly among 13-year olds. Aside from the concern of general decline, then, there is evidence of increasing disparity in the relative positions of communities as time progresses.

CHAPTER IV

CONCLUSIONS

The Reiser group-effects model was successfully used to link data across two age levels and over two time points in each of three skill areas of the National Assessment of Educational Progress surveys of mathematics. Experience gained in this effort lead to several important conclusions concerning the application of item response methods in general and of the group-effects model in particular to the National Assessment.

Items grouped at the level of NAEP subtopics proved satisfactory for scaling with a unidimensional model, even across age levels and assessment years. Goodness-of-fit indices within the age/year data matrices and successful links across ages and years imply that trends and group differences can be profitably analyzed at this higher level of abstraction than the individual item, yet allowing for the administration of different subsets of items to different age groups and at different points in time. This finding is particularly fortuitous when seen in the light of the NAEP multiple-matrix sampling design; the items from a subtopic are generally spread over several test booklets. Such a scheme yields more precise estimates of group-level attainment than a scheme that presented more items from a scale to fewer different persons.

More inclusive and broader-ranged collections of items would not have lead to satisfactory results. The combined calibration of Arithmetic Computation and Understanding Mathematical Concepts, for example, could not have shown how blacks were closing the gap from whites in the former area but lagging further behind in the latter. The need for scales that maintain their integrity over time, then, requires rather narrow domains for scaling. While it may be convenient with current NAEP data tapes to scale together all the items that happen to appear in the same booklet, the intentional heterogeneity of such a collection virtually guarantees a poor fit to any unidimensional item response model and severe item parameter drift over time. Under current NAEP item-sampling designs, the practice of item response scaling within NAEP booklets should be most strongly discouraged.

Given that scaling must be accomplished within fairly narrow skill area (e.g., NAEP subtopics), methods of summarizing results over these areas must be determined. If levels of performance increase in computational skills but decrease in understanding concepts, as an example, what should be said about skill in mathematics as a whole? Clearly some scheme of indexing or weighted averaging is required, with explicit rules by which the information from the separate skills is combined.

Within these restrictions, alternative methods of scaling are available. This project has made more clear some of the advantages and disadvantages of one of those alternatives, namely, the Reiser model for group effects.

Of great advantage to this project was the fact that numbers of attempts and correct responses to each item in a scale from each cell in a design on persons are sufficient for estimating item parameters and group effects. A summary file at the level of groups of persons rather than a full file at the level of individuals need be handled. This same feature of the model, however, may be seen as a disadvantage as well. Because the model adresses data at the level of cells in the design on persons, there are practical limits to the complexity of the design that may be employed before the numbers of persons in the cells become too small. The design used in these examples contained sex, race/ethnicity, region of the country, and size and type of community--168 cells in all. Several of these cells were small or empty, and it is clear that not many additional factors could be included in the design before there were more cells than observations.

In sum, these applications of the group-effects model can be considered successful as a demonstration of the practicality of applying item-response methods to the efficient multiple-matrix data of modern assessments. Whether the group-effects model or a close cousin eventually dominates, the generic advantages of item response theory are sure to advance the practice of assessment.

REFERENCES

Anastasi, A. Differential Psychology. New York: Macmillin, 1958.

Andersen E.B. & Madsen M. Estimating the parameters of a latent population distribution. Psychometrika, 1977, 42, 357-374.

Bock, R.D. Multivariate Statistical Methods in Behavioral Research. New York, NY: McGraw-Hill, 1975.

Bock, R.D. Basic issues in the measurement of change. In D.N.M. de Gruijter & L.J.T. van der Kamp (Eds.), Advances in Educational and Psychological Measurement. London: John Wiley & Sons, 1976.

Bock, R.D. A feasibility study of the one-, two-, and three-parameter logistic response models for the analysis and reporting of California Assessment data. Chicago: International Educational Services, 1979.

Bock, R.D., & Mislevy, R.J. An item response-curve model for matrix-sampling data: The California grade-3 assessment. New Directions for Testing and Measurement. Number 10, 1981.

Cox, D.R. The Analysis of Binary Data. London: Methuen, 1970.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements. New York: Wiley, 1972.

Haebera, T. Broad-range item calibration by constrained optimization of item parameter transformation. Paper presented at the meeting of the Psychometric Society, Chapel Hill, NC, May 1981.

Lord, F.M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord F.M. & Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Mislevy, R.J. Efficiency, objectivity, and robustness in trait estimation. Paper presented at the meeting of the Psychometric Society, Chapel Hill, NC, May 1981.

Mislevy, R.J. Estimating group-level attainment from sparse multiple-matrix samples of item responses. Paper presented at the meeting of the American Educational Research Association, New York, New York, March 1982.

Pandey, T. & Carlson, D. Assessing payoffs in the estimation of the mean using multiple matrix sampling designs. In D.N.M. de Gruijter & L.J.T. van der Kamp (Eds.), Advances in Educational and Psychological Measurement. London: John Wiley & Sons, 1976.

Reiser, M.R. A Latent Trait Model for Group-Effects. Unpublished Ph.D. Dissertation. Department of Behavioral Sciences, University of Chicago, 1980.

Sanathanan, L. and Blumenthal, N. The logistic model and estimation of latent structure. Journal of the American Statistical Association, 1978, 73, 794-798.

Tucker, L. A method for scaling ability test items in difficulty taking item unreliability into account. American Psychologist, 1948, 3, 309.

Tyler, R.W. What is an Ideal Assessment Program? Sacramento, Bureau of Research Services, California State Department of Education, 1968.

Womer, F.B. Developing a Large Scale Assessment Program. Denver: Colorado Department of Education, 1973.

APPENDIX A

A LATENT TRAIT MODEL FOR GROUP EFFECTS

A Detailed Specification of the Model

The development of the first two sections of this chapter parallels that of Bock (1976). In the first part the model is stated in terms of a binomial response function. In the second part, maximum likelihood estimates are derived for the parameters in the model. The last section consists of a discussion of the asymptotic properties of the estimates. A test of fit for the model is also discussed in this section.

Two symbols which are used repeatedly in this chapter require a brief explanation. S is used as a summation sign, instead of the more common upper case sigma, and d is used as the symbol indicating a derivative.

Assume that subjects respond to one item from the set of items which constitute the scale, and that subjects are assigned to f homogeneous sample groups. Assume also that subjects in group q are a probability sample from a conditionally normal latent trait distribution with mean represented by the contrast $\underline{k}_q{}'\underline{\theta}$ and variance σ^2.

$\underline{k}_q{}'$ represents the q^{th} row of the general design matrix K,

$$K = \left|\begin{array}{ccccc} k_{11} & k_{12} & k_{13} & \cdots & k_{1s} \\ k_{21} & & & & \\ k_{31} & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ k_{f1} & \cdot & \cdot & \cdot & k_{fs} \end{array}\right|$$

s is the rank of the model for estimation.

$\underline{\Theta}$ represents a vector of contrasts among the group effects.

A subject's response to item j is scored

$$h_{qj} = \begin{cases} 1 \text{ if correct} \\ 0 \text{ otherwise} \end{cases}$$

The probability that the subject (or respondent) responds correctly is given by the logistic ogive (logit) model. The logistic curve is used here as an approximation to the much more complicated mormal cumulative distribution function. Haberman (1974, pg 34) concludes that no emperical evidence exists that the normal distribution provides more accurate models than the logistic. So,

(1) $P(h_{qj} = 1) = F(z_{qj}) = 1/(1 + \exp(-z_{qj}))$, and

$P(h_{qj} = 0) = 1 - F(z_{qj})$

As mentioned in chapter 1, the design of the sample groups is introduced into the specification of the logit, z_{qj}.

$$z_{qj} = c_j + \underline{a}_j k_q{}' \underline{\Theta},$$

where c_j and a_j are parameters for item j.

The priniple OF local independence states that responses to items are independent, conditional on item and group parameters. By this principle, the probability of r_{qj} correct responses from the N_{qj} respondents in group q who attempt item j is given by the binomial function:

$$P(r_{qj} | N_{qj}, \underline{k}_q{}' \underline{\Theta}, c_j, a_j) = \frac{N_{qj}!}{r_{qj}!(N_{qj} - r_{qj}!)} F^{r_{qj}}(z_{qj})[1 - F(z_{qj})]^{N_{qj} - r_{qj}}$$

The probability of the entire sample is taken over groups and items:

$$(2) \qquad P = \prod_{q}^{f} \prod_{j}^{n} P(r_{qj} | N_{qj}, \underline{k}_q{}' \underline{\Theta}, \underline{c}_j, \underline{a}_j)$$

For a t-way design on the subjects, the factor indices i_w vary from 1 to m_w for $w = 1,...t$. The number of cells in the design, f, is equal to ${}_{w} m_w$. Any cell in the design can be referenced by a single subscript q as follows:

$$q = i_1 + \underset{2}{S} (i_w - 1) \prod_{r<w} m_r$$

Equation (1) specifies a quantal-response type model which is closely related to the probit model of Finney (1971) and the logit model of Berkson (1944). Finney uses the cumulative normal distribution, but as stated before, there is no emperical evidence for prefering the normal over the logistic, and the logistic is considerably simpler. The logistic quantal response models are log-linear models, and thus many of the methods and results from Haberman (1974) are applicable to the present model.

Derivation of Parameter Estimates.

Estimates for the item and group parameters can be obtained in a straight forward manner by the method of maximum likelihood. As will be seen in this section, there are two linear dependencies among the set of vectors which consists of the columns of the information matrix. In order to eliminate these dependencies, two parameters must be fixed arbitrarily. One choice which can be made here would be to fix the first and the m_1th effects from the first factor of the design respectively. This sets the scale of all the estimates in a very convenient range. Another choice for eliminating one of the dependencies would be to include a prior distribution for the item slope parameter within the model. Some previous experience with two parameter models has shown that including this prior knowledge results in a more well behaved solution in the sense that parameter estimates for items on which there is little information in the data will not take on a value which is unduly large. What happens in practice is that the slope parameter can become very high for an item on which the responses are either nearly all correct or nearly all incorrect. For such an item, the information provided by the prior distribution becomes dominant, and the solution is primarily a function of this information.

Since such items add essentially nothing to the likelihood of the data, eliminating them from the analysis entirely constitutes an equally effective strategy. However, the prior distribution alternative renders the model more robust in the sense that less work with the data will be necessary before satisfactory estimates are obtained. Consequently, at some points in the derivation, information will be included describing changes that would be required in the equations in order to obtain maximum aposteriori density (MAP) estimates. A complete derivation of the MAP estimates would be nearly the same as the derivation of the maximum likelihood estimates, and so it would be needlessly repetitive.

For the maximum aposteriori density estimates, the slope parameter, a_j, is assumed to be distributed log normally with mean μ_a and variance σ_a^2. These two parameters for the distribution are given values by the researcher before estimates of the item parameters are obtained. A state of nearly total ignorance about the prior distribution can be indicated by specifying a large variance. The results of Lindley and Smith (1972) show that it is more reasonable to estimate the mode rather the the mean of the posterior distribution, so the easier path will be taken here.

For maximum likelihood estimates, the likelihood of the entire sample is obtained directly from expression (2):

$$\ell(\underline{c},\ \underline{a},\ \underline{e}) = \underset{q}{s}\ \underset{j}{s}\ [\text{const} + r_{qj} \log F(z_{qj}) +$$

$$(N_{qj} - r_{qj}) \log(1 - F(z_{qj}))]$$

Bock and Thissen (1979) show the general form for the logarithm of the posterior density. In this setting, it takes the following form:

$$\lambda(\underline{c},\ \underline{a},\ \underline{e}) = \underset{q}{s}\ \underset{j}{s}\ [\text{const} + r_{qj} \log F(z_{qj}) +$$

$$(N_{qj} - r_{qj}) \log(1 - F(z_{qj}))] - \frac{1}{2}\, s\, \frac{(\log a_j - \mu_a)^2}{\sigma_a^{\ 2}}$$

Notice that the difference betweeen these two equations consists only of a term, after the minus sign on the right, which represents the prior information.

The following are obtained now for use later:

$$F(z_{qj}) = \exp(z_{qj})/(\exp(z_{qj}) + 1) = 1/(1 + \exp(-z_{qj})$$

$$1 - F(z_{qj}) = \exp(z_{qj}) + 1)/(\exp(z_{qj}) + 1) \quad -$$

$$\exp(z_{qj})/(\exp(z_{qj}) + 1)$$

$$= 1/(1 + \exp(z_{qj})) = \exp(-z_{qj})/(\exp(-z_{qj}) + 1$$

$$\frac{dF(z_{qj})}{dz_{qj}} = -1(1 + \exp(-z_{qj}))^{-2}\, \frac{d \exp(-z_{qj})}{dz_{qj}}$$

$$= \exp(-z_{qj})/(1 + \exp(-z_{qj}))(1 + \exp(-z_{qj}))$$

$$= F(z_{qj})[1 - F(z_{qj})]$$

$$\frac{d[1 - F(z_{qj})]}{dz_{qj}} = -F(z_{qj})[1 - F(z_{qj})$$

$$\frac{dz_{qj}}{dc_j} = 1 \qquad \frac{dz_{qj}}{da_j} = \underline{k}_q{}'\underline{\Theta} \qquad \frac{dz_{qj}}{d\Theta_g} = k_{qg}a_j$$

Once the likelihood function has been chosen, the maximum of the functin with respect to a given parameter is often the point at which the rate of change of the function, the first derivative, is equal to zero. Such a point could also be a minimum or a boundary point, so other aspects of the likelihood function have to be investigated. We will attend to these other aspects shortly.

Maxima:

$$\frac{d\ell}{dc_j} = S\ [\ \frac{r_{qj}}{F(z_{qj})} F(z_{qj})\ [1 - F(z_{qj})]\ \frac{dz_{qj}}{dc_j} \qquad +$$

$$\frac{(N_{qj} - r_{qj})}{[1 - F(z_{qj})]} (-1)\ F(z_{qj})[1 - F(z_{qj})]\ \frac{dz_{qj}}{dc_j}]$$

$$= \underset{q}{S}\ [r_{qj} - N_{qj}F(z_{qj})] = 0$$

$$\frac{d\ell}{da_j} = S\ [r_{qj} - N_{qj}F(z_{qj})]\ \underline{k}_q{}'\underline{\theta} = 0$$

$$\frac{d\ell}{d\theta_g} = \underset{q}{S}\ \underset{j}{S}\ [r_{qj} - N_{qj}F(z_{qj})]\ k_{qg}a_j = 0$$

For MAP estimates, the preceeding equations would differ only in the presence of the so-called penalty function as a second term in derivative with respect to a_j:

$$\frac{d\ell}{da_j} = S\ [r_{qj} - N_{qj}F(z_{qj})]\ \underline{k}_q{}'\underline{\theta} - \frac{\log a_j - \mu_a}{a_j^2\sigma_a^2} = 0$$

If B_{qj} is set equal to $r_{qj} - N_{qj}F(z_{qj})$, the preceeding likelihood equations can be rewritten in simpler expressions:

$$c_j:\ \underset{q}{S}\ B_{qj} = 0$$

$$a_j:\ \underset{q}{S}\ B_{qj}\ \underline{k}_q{}'\underline{\theta} = 0$$

$$\theta_g:\ \underset{q}{S}\ \underset{j}{S}\ B_{qj}\ k_{qg}a_j = 0$$

These equations cannot be solved explicitly for the unknown parameters, but estmates can be obtained by an iterative numerical procedure such as that of Newton-Raphson. Second derivatives of the log likelihood are required for the purpose of investigating the shape of the likelihood function and for use in the Newton-Raphson procedure.

Second derivatives:

$$\frac{d^2 \ell}{dc_j\, dc_h} = \delta_{jn} \underset{q}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]$$

$$\frac{d^2 \ell}{dc_j\, da_h} = \delta_{jn} \underset{q}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]\ \underline{k}_q{}'\underline{\Theta}$$

$$\frac{d^2 \ell}{da_j\, da_h} = \delta_{jh} \underset{q}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]\ (\underline{k}_q{}'\underline{\Theta})^2$$

$$\frac{d^2 \ell}{da_j\, d\Theta_g} = \underset{Q}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]\ k_{qg}a_j$$

$$\frac{d^2 \ell}{da_j\, d\Theta_g} = \underset{q}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]\ (\underline{k}_q{}'\underline{\Theta})k_{qg}a_j + B_{qj}k_{qg}$$

$$\frac{d^2 \ell}{d\Theta_g\, d\Theta_h} = \underset{q}{S}\ \underset{j}{S} -N_{qj}F(z_{qj})[1 - F(z_{qj})]\ k_{qg}a_jk_{qh}a_j$$

where $\delta_{jn} = \begin{cases} 1 \text{ if } j=n \\ 0 \text{ otherwise} \end{cases}$ is known as Kronecker's delta.

Only one of the second derivatives would differ if we were deriving MAP estimates. The derivative taken twice with respect to the slope would include another term on the right hand side of the equation:

$$\frac{d^2 \ell}{da_j \; da_h} = {}_{jh} \underset{q}{S} -N_{qj} F(z_{qj}) [1 - F(z_{qj})] (\underline{k}_q{}' \underline{\theta})^2 - \frac{1 - \log a_j + \mu_a}{\sigma_a^2 \; a_j^2}$$

By using the expected values for r_{qj} and B_{qj} in the above equations the elements of the information matrix can be obtained.

$$E(r_{qj}) = N_{qj} F(z_{qj})$$

$$E(B_{qj}) = N_{qj} F(z_{qj}) - N_{qj} F(z_{qj}) = 0$$

Also, set $W_{qj} = F(z_{qj}) [1 - F(z_{qj})]$.
The elements of the information matrix are then as follows:

$$A_{11}: \quad E\left(- \frac{d^2 \ell}{dc_j \; dc_h}\right) = \delta_{jh} \underset{q}{S} N_{qj} W_{qj}$$

$$A_{21}: \quad E\left(- \frac{d^2 \ell}{dc_j \; da_h}\right) = \delta_{jh} \underset{q}{S} N_{qj} W_{qj} \; (\underline{k}_q{}' \underline{\theta})$$

$$A_{22}: \quad E\left(- \frac{d^2 \ell}{da_j \; da_h}\right) = \delta_{jh} \underset{q}{S} N_{qj} W_{qj} (\underline{k}_q{}' \underline{\theta})^2$$

$$B_1: \quad E\left(-\frac{d^2 \ell}{dc_j \, d\Theta_g}\right) = \underset{q}{S} \, N_{qj} W_{qj} k_{qg} a_j$$

$$B_2: \quad E\left(-\frac{d^2 \ell}{da_j \, d\Theta_g}\right) = \underset{q}{S} \, N_{qj} W_{qj} (\underline{k}_q{}' \, \underline{\Theta}) k_{qg} a_j$$

$$C: \quad E\left(-\frac{d^2 \ell}{d\Theta_g \, d\Theta_h}\right) = \underset{q}{S} \, \underset{j}{S} \, N_{qj} W_{qj} k_{qg} a_j k_{qh} a_j$$

The information matrix, $I(\underline{c}, \underline{a}, \underline{\Theta})$, takes the form

$$I(\underline{c}, \underline{a}, \underline{\Theta}) = X'WX \qquad \text{where}$$

$$W = \text{diag}[N_{11}W_{11}, N_{21}W_{21}, \ldots N_{f1}W_{f1}, N_{12}W_{12}, \ldots N_{fn}W_{fn}]$$

is positive definite.

of linear combinations of columns with linearly independent coefficients 0, 0, 0, . . . 0, a_1, a_2, a_3, . . . a_n, $-\theta_1$, $-\theta_2$, $-\theta_3$, . . . $-\theta_s$. It is anticipated that identity contrasts will always be used over the first factor during parameter estimation, hence another dependency exists among the columns of X as a result of these m_1 identity contrasts. The linear combination of columns with linearly independent coefficients a_1, a_2, a_3, . . . a_n, 0, 0, 0, . . . 0, -1, -1, -1, . . . -1, 0, 0, 0, . . . 0 shows the dependence. If any two parameters are arbitrarily fixed and the corresponding likelihood equations deleted, the information matrix with the corresponding rows and columns deleted is positive definite. A necessary and sufficient condition for the log-likelihood function to be concave and have a unique maximum is that this Hessian matrix (matrix of negative of expected value of second derivatives) is positive definite. In the limit, therefore, unique maximum likelihood estimates of the parameters exist.

The information matrix can be written in partitioned form:

$$I(\underline{c}, \underline{a}, \underline{\theta}) = \begin{vmatrix} A & B' \\ B & C \end{vmatrix}$$

where

$$A = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}$$

$$A_{11} = \text{diag}(\underset{q}{S} N_{qj} W_{qj})$$

of linear combinations of columns with linearly independent coefficients 0, 0, 0, . . . 0, a_1, a_2, a_3, . . . a_n, $-\theta_1$, $-\theta_2$, $-\theta_3$, . . . $-\theta_s$. It is anticipated that identity contrasts will always be used over the first factor during parameter estimation, hence another dependency exists among the columns of X as a result of these m_1 identity contrasts. The linear combination of columns with linearly independent coefficients a_1, a_2, a_3, . . . a_n, 0, 0, 0, . . . 0, -1, -1, -1, . . . -1, 0, 0, 0, . . . 0 shows the dependence. If any two parameters are arbitrarily fixed and the corresponding likelihood equations deleted, the information matrix with the corresponding rows and columns deleted is positive definite. A necessary and sufficient condition for the log-likelihood function to be concave and have a unique maximum is that this Hessian matrix (matrix of negative of expected value of second derivatives) is positive definite. In the limit, therefore, unique maximum likelihood estimates of the parameters exist.

The information matrix can be written in partitioned form:

$$I(\underline{c}, \underline{a}, \underline{\theta}) = \begin{vmatrix} A & B' \\ B & C \end{vmatrix}$$

where

$$A = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}$$

$$A_{11} = \mathrm{diag}(\underset{q}{S} N_{qj} W_{qj})$$

$$A_{21} = A_{12} = \text{diag}\,(\underset{q}{S} N_{qj} W_{qj} (\underline{k}_q{}' \underline{\Theta}))$$

$$A_{22} = \text{diag}\,(S\, N_{qj} W_{qj} (\underline{k}_q{}' \underline{\Theta})^2)$$

$$B = \left| B_1 \quad B_2 \right| = [\underset{q}{S} N_{qj} W_{qj} k_{qg} a_j, \underset{q}{S} N_{qj} W_{qj} (\underline{k}_q{}' \underline{\Theta}) k_{qg} a_j]$$

$$C = \left| S\, S\, N_{qj} W_{qj} k_{qg} a_j k_{qh} a_j \right|$$

Since X has a deficiency in rank of 2, two parameters can be arbitrarily fixed and the corresponding likelihood equations deleted. As discussed earlier in this chapter, the first and m_1th effects from the first factor of the design are the parameters chosen to be fixed at -1 and +1 respectively. This choice conveniently sets the scale of the solution in terms of the range of the first factor effects. As also discussed previously, one of the linear dependencies can be eliminated by specifying a prior distribution on the slope paramters instead of arbitrarily fixing a second parameter. The dependency eliminated by the prior would be the one associated with the linearly independent coefficients $0, 0, 0, \ldots 0, a_1, a_2, a_3, \ldots a_n, -\Theta_1, -\Theta_2, -\Theta_3, \ldots -\Theta_s$.

The inclusion of the prior distribution does not change the composition of the X matrix, but the X matrix is never actually formed during the estimation procedure. The information matrix is formed in the the three partitions, A, B, and C. The last n columns of the submatrix A are formed by the inner products of the $n^{th} + 1$ through $2n^{th}$ columns of the X matrix. These columns are linearly dependent on the last s columns of X. Now the prior distribution is included by adding the matrix G, say, where

$$G = \text{diag}\left(0, 0, 0, \ldots 0, \frac{1 - \log a_1 + \mu_a}{\sigma_a^2 a_1^2}, \frac{1 - \log a_2 + \mu_a}{\sigma_a^2 a_2^2}, \ldots \right.$$

$$\left. \frac{1 - \log a_n + \mu_a}{\sigma_a^2 a_1^2}, 0, 0, 0, \ldots 0\right)$$

to the informtion matrix, which has the effect of adding a term to each of the last n diagonal elements of A, A being 2n by 2n. The linear dependency among the last columns of A and the other rows (columns) of the information matrix is thus eliminated by the addition of the elements of G to the diagonal, and the additional row and column need not be deleted in this case.

For the model with no prior distribution on the slope parameters, two rows and columns corresponding to two group effects are deleted, and the information matrix will be positive definite. Then,

$$I(\underline{c}, \underline{a}, \underline{\theta}) = \begin{bmatrix} A & B^{*'} \\ B^{*} & C^{*} \end{bmatrix}$$ is the $2n + s - 2$

rank information matrix. If a prior distribution is specified for the slopes, rows and columns corresponding to only one effect are deleted. $I(\underline{c}, \underline{a}, \underline{\theta}^*)$ will still be positive definite, but the rank will be $2n + s - 1$.

For the MAP estimation, the information matrix is adjusted when used in the Newton-Raphson iterations for the influence of the prior distribution, resulting in the matrix, say, E.

$$E = I(\underline{c}, \underline{a}, \underline{\theta}^*) + G$$

where G takes the form as defined previously.

For the regular maximum likelihood estimates, no adjustment is made to the information matrix.

Proceeding to obtain the necessary quantities for the scoring solution, we need the inverse of the information matrix, or the information matrix as adjusted for the prior distribution.

$$I^{-1} =$$

$$\begin{bmatrix} A^{-1} + A^{-1}B^{*\prime}(C^* - B^*A^{-1}B^{*\prime})^{-1}B^*A^{-1} & -A^{-1}B^{*\prime}(C^* - B^*A^{-1}B^{*\prime})^{-1} \\ -(C^* - B^*A^{-1}B^{*\prime})^{-1}B^*A^{-1} & (C^* - B^*A^{-1}B^{*\prime})^{-1} \end{bmatrix}$$

There are some aspects of I^{-1} which can be used for efficient computing. The whole matrix is of course Grammian, so the upper right partition is simply the transpose of the lower left partition. The right hand term in the upper left partition, $A^{-1}B^{*'}(C - B^{*}A^{-1}B^{*'})^{-1}B^{*}A^{-1}$, is Grammian, and can be formed with specialized routines from the two matrices $B^{*}A^{-1}$ and $(C^{*} - B^{*}A^{-1}B^{*'})^{-1}$, which is the lower right partition. The matrix a^{-1} does not require heavy computation because the matrix a consists of partitions which are diagonal:

$$A^{-1} = \begin{vmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{vmatrix}$$

where

$$A^{11} = D^{-1}A_{22} \qquad A^{21} = A^{12} = -D^{-1}A_{12} \qquad A^{22} = D^{-1}A_{11}$$

$$D = \text{diag}[(\underset{q}{S} N_{qj}W_{qj})(\underset{q}{S} N_{qj}W_{qj}(\underline{k}_q{}'\underline{G})^2) - (\underset{q}{S} N_{qj}W_{qj}(\underline{k}_q{}'\underline{\Theta})^2]$$

The largest matrix to be directly inverted is the s - 2 rank $(C^{*} - B^{*}A^{-1}B^{*'})$.

For the MAP estimates, I^{-1} is replaced by E^{-1}. E^{-1} is the same as I^{-1} except for the contents of A^{11} and D. So, if the prior distribution is specified on the slopes, A^{11} and D become as follows:

$$A^{11} = D^{-1}A_{22} + \text{diag}\left(\frac{1 - \log a_1 + \mu_a}{\sigma_a^2 \, a_1^2}, \frac{1 - \log a_2 + \mu_a}{\sigma_a^2 \, a_2^2}, \ldots, \frac{1 - \log a_3 + \mu_a}{\sigma_a^2 \, a_n^2}\right)]$$

103

$$D = \mathrm{diag}[(\underset{q}{S} N_{qj}W_{qj})(\underset{q}{S} N_{qj}W_{qj}(\underline{k}_q{}'\underline{\theta})^2 + \frac{1 - \log a_j + \mu_a}{\sigma_a{}^2\, a_j{}^2}) - (\underset{q}{S} N_{qj}W_{qj}(\underline{k}_q{}'\underline{\theta})^2)]$$

Here, $(C^* - B^*A^{-1}B^{*\prime})$ is of rank s - 1.

The Newton-raphson procedure consists of finding estimates at the t+1th iteration by adding a correction to the estimates at the tth iteration. The correction is obtained from multiplying the inverse of the matrix of second derivatives by the matrix of first derivatives. If the information matrix, which contains expected values for the second derivatives, is substituted for the matrix of actual second derivatives, the iterative proceedure with this substitution is known as Fisher's method of Efficient Score:

$$\begin{vmatrix} \underline{c} \\ \\ \underline{a} \\ \\ \underline{\theta}^* \end{vmatrix} \qquad \begin{vmatrix} \underline{c} \\ \\ \underline{a} \\ \\ \underline{\theta}^* \end{vmatrix} + I_t(\underline{c}, \underline{a}, \underline{\theta}^*)^{-1} \begin{vmatrix} \underset{q}{S} B_{qj} \\ \\ \underset{q}{S} B_{qj}(\underline{k}_q{}'\underline{\theta}) \\ \\ \underset{q}{S}\underset{j}{S} B_{qj}k_{qj}a_j \end{vmatrix}$$

In extremum theory, the vector of first derivatives is known as the gradient. The maximum of the likelihood function exists at the zero of the gradient. The second derivatives tell how fast the gradient is changing. As the gradient approaches its zero, it will change faster and faster, and the elements of the inverse of the information matrix will become smaller and smaller. So, at the maximum of the likelihood function, the correction to be added to the estimates becomes zero. The iterative process is stopped and considered converged whenever the absolute value for all corrections falls below a preassigned criterion. Starting values of 0s and 1s, for the c_j's and a_j's respectively, have been used with success. Least squares estimates of the group effects calculated on the cell proportions can be used as starting values for $\underline{\Theta}^*$.

Asymptotic Properties

Many of the traditional results which hold for maximum likelihood estimates are useful here. Since the only parameters associated with the subjects are fixed group effects, this model avoids one of the thorniest problems often encountered by two parameter latent trait models. In such a model where each subject has an ability to be estimated, the subject parameter, which appears as a niusance parameter, cannot be conditioned out of the likelihood equations, and the number of parameters increases with the number of respondents. In the present model, however, the number of parameters are fixed even as the number of respondents becomes very large. Hence, standard results that are covered in general treatments such as

Cramer (1946) and Rao (1965) apply.

Maximum likelihood estimates have the properties of consistency and asymptotic efficiency, the latter meaning that the variance of the estimates is the minimum attainable by any consistent estimator. Additionally, the estimates are distributed in multivariate normal form, with variance-covariance matrix equal to the inverse of the negative of the matrix of second derivatives, i.e., the information matrix. This information measure, also known as Fisher's information, proves to be a general index of sensitivity for small changes in the value of the parameter (Rao, 1962).

The standard errors for the estimates are formed from elements of the information matrix as follows:

$$\text{S.E.}(c_j) = 1/\text{SQRT}\left(I^{-1}_{jj}\right)$$

$$\text{S.E.}(a_j) = 1/\text{SQRT}\left(I^{-1}_{n+j,\,n+j}\right)$$

$$\text{S.E.}(\Theta_g) = 1/\text{SQRT}\left(I^{-1}_{2n+j,\,2n+j}\right)$$

Fortunately, the terms needed for the denominators of the expressions can be taken directly from the information matrix as formed during the last iteration of the scoring procedure.

Testing Goodness of Fit

Two statistics which are commonly used as a measure of the distance between the model and the data are the likelihood ratio chi-square, sometimes written as G^2, and the Pearson chi-square, sometimes written as X^2. G^2 and X^2 are defined as follows:

$$G^2 = 2 \sum_{j}^{n} \sum_{q}^{f} \sum_{h}^{2} r_{qjh} \log \frac{r_{qjh}}{N_{qj} P_{qjh}}$$

where h is over all responses to an item.

$$X^2 = \sum_{j}^{n} \sum_{q}^{f} \frac{(N_{qj} - N_{qj} P_{qj})^2}{N_{qj} P_{qj}}$$

The degrees of freedom for these statistics are equal to nf - 2n - s + 2. In practice, the values of G^2 and X^2 are essentially the same for a given model and data set, although G^2 may be more resistent to ill effects of cells with very low expected value. It can be shown quite readily that X^2 is a sum of squares of approximate unit normal deviates, and has, therefore, approximately a chi-square distribution on nf - 2n - s + 2 degrees of freedeom (see for example, Brownlee (1965)). Bishop, Fienberg, and Holland (1975) show that G^2 and X^2 are asymptotically equivalent under the correct model for the data. G^2 has the overwhelming advantage that it can be used for comparing alternative nested models using a conditional breakdown of the chi-square measures for the models.

APPENDIX B

A MINIMUM CHI-SQUARE SOLUTION FOR LINKING CALIBRATIONS FROM TEST FORMS WITH AN ARBITRARY DESIGN OF OVERLAPPING

A WEIGHTED LEAST-SQUARES SOLUTION FOR LINKING CALIBRATIONS FROM FORMS WITH AN ARBITRARY DESIGN OF OVERLAP

Robert J. Mislevy

International Educational Services

INTRODUCTION

The 2-parameter logistic item response model expresses the probability of a correct response to Item j from Agent i as

$$P_{ij} = \Psi \left[\frac{\theta_i - \beta_j}{\sigma_j} \right], \qquad (1)$$

where

$\Psi(x)$ denotes the logistic function $\exp(x)/[1+\exp(x)]$,

θ_i is the ability parameter for Agent i,

β_j is the threshold parameter of Item j, and

σ_j is the dispersion parameter of Item j (the recip- of the slope parameter of Item j).

Reiser's (1980) latent trait model for group effects follows this form, with "Agent i" interpreted as the group of subjects in a specified cell of the NAEP demographic sampling design, and with θ_i being a linear function of a vector of group-effect parameters.

Item and group parameters are determined uniquely only up to a linear transformation. When subsets of items from the same scale are calibrated in separate data sets (e.g., data from different assessment years or different age groups), linear transformations must be found which optimally rescale item and group-parameter estimates from any given calibration to a common scale with a specified origin and unit-size.

The method of linking calibrations described in this paper is intended for the case in which two or more calibration runs have been performed on independent sets of data. In each case, all items are assumed to belong to the same scale. It is necessary that each calibration contain at least two items that appear in some other calibration, and that all calibrations are linked either directly or indirectly. (Calibrations k and ℓ are linked directly if they have items in common; they are linked indirectly if Calibration j shares items with Calibration h, which in turn shares items with Calibration ℓ. Any such chain of finite length constitutes an indirect link.) The method utilizes information from all links among all calibrations in the estimation of optimal transformations to a common scale.

SETTING UP NOTATION

We concern ourselves with item and group parameter estimates from M separate calibrations. Item parameter estimates are denoted as follows:

Bjk is the estimate of the threshold parameter of Item j from Calibration k, if Item j has been included in Calbration k; otherwise, this value is undefined;

Sjk is the estimate of the dispersion parameter of Item j from Calibration k, if Item j has been included in that calibration run.

θik is the estimate of the ability of Group i obtained in Calibration k, if appropriate.

The linear transformations we seek will, for convenience, rescale the estimates from all other calibrations to the scale determined in Calibration 1. They are denoted as follows:

$$Lk(x) = Ak\ x + Ck.$$

They are applied to the estimates as follows:

$$\theta ik^* = Ak\ \theta ik^* + Ck,$$

$$Bjk^* = Ak\ Bjk^* + Ck, \text{ and}$$

$$Sjk^* = Ak\ Sjk.$$

It is clear that each item will have at least two estimates of each of its parameters, after these transformations have been applied to the results from each calibration. Inasmuch as the transformations represent optimal rescaling to a common unit and origin, final estimates of item parameters may be obtained by taking the averages of the estimates for a particular value, with each estimate weighted by the squared reciprical of its rescaled standard error of estimation.

THE FITTING FUNCTION

The weighted least-squares fitting function that simultaneously estimates the transformations for Calibrations 2 through M, using information from all available links, is shown below. It is to be understood that A_1 is fixed at 1 and C_1 at 0.

$$F = \sum_{j=1}^{N} \sum_{k=1}^{M} \sum_{\ell>k}^{M} \left\{ \left[(A_k B_{jk} + C_k) - (A_\ell B_{j\ell} + C_\ell) \right]^2 \cdot W_{jk\ell} + \left[(A_k S_{jk} - A_\ell S_{j\ell})^2 \right] W^*_{jk\ell} \right\}$$

where

$$W_{jk\ell} = \begin{cases} \{\mathrm{Sqrt}[A_k^2\, SE^2(B_{jk}) + A_\ell^2\, SE^2(B_{j\ell})]\}^{-1} & \text{if Item } j \text{ is included in both calibrations,} \\ 0 & \text{otherwise;} \end{cases}$$

$$W^*_{jk\ell} = \begin{cases} \{\mathrm{Sqrt}[A_k^2\, SE^2(S_{jk}) + A_\ell^2\, SE^2(S_{j\ell})]\}^{-1} & \text{if Item } j \text{ is included in both calibrations,} \\ 0 & \text{otherwise.} \end{cases}$$

A computational method for obtaining the minimum of the fitting function may begin with an unweighted least-squares approximation which uses information from threshold estimates only, as described in the following section. In this section, we provide approximate first and second derivatives of the fitting function with respect to the parameters of the transformations, which may be used in a quasi-Newton solution. Given approximations $\underline{A}^{(m)}$ and $\underline{C}^{(m)}$ of the parameters of the transformations, we obtain better approximations as follows:

$$\begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}^{(m+1)} = \begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}^{(m)} - \left.\begin{bmatrix} \frac{d^2F}{d\underline{A}\,d\underline{A}} & \frac{d^2F}{d\underline{A}\,d\underline{C}} \\ \frac{d^2F}{d\underline{C}\,d\underline{A}} & \frac{d^2F}{d\underline{C}\,d\underline{C}} \end{bmatrix}^{-1}\right|_{\begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}=\begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}^{(m)}} \left.\begin{bmatrix} \frac{dF}{d\underline{A}} \\ \frac{dF}{d\underline{C}} \end{bmatrix}\right|_{\begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}=\begin{bmatrix}\underline{A}\\ \underline{C}\end{bmatrix}^{(m)}}$$

The presence of the slope parameters Ak in the weights complicates the computation of derivatives. We propose, therefore, that during the iterative solution of this problem, the weights be considered as constants at each step. That is, during the computation of the (m+1)'th estimates, the weights are to be computed from the known values of the standard errors of the item parameter estimates and the transformation slope parameter estimates Ak obtained from the m'th step. This expedient can be expected to have little effect on the efficiency of the solution. Under this assumption, we obtain the first and second derivatives of the fitting function F as shown below. It is to be understood that these derivatives are for transformations 2 through M.

First derivatives

$$\text{Ak:}\quad 2 \sum_{j=1}^{N} \sum_{\substack{\ell=1 \\ (\ell\neq k)}}^{M} \Big[(A_k B_{jk}^{2} - A_\ell B_{jk} B_{j\ell} + C_k B_{jk} - C_\ell B_{j\ell})\, W_{jk\ell} + (A_k S_{jk}^{2} - A_\ell S_{jk} S_{j\ell})\, W^{*}_{jk\ell} \Big]$$

$$\text{Ck:}\quad 2 \sum_{j=1}^{N} \sum_{\substack{\ell=1 \\ (\ell\neq k)}}^{M} (A_k B_{jk} - A_\ell B_{j\ell} + C_k - C_\ell)\, W_{jk\ell}$$

Second derivatives

$$\text{Ak, Ak:}\quad 2 \sum_{j=1}^{N} \left\{ \Big[B_{jk}^{2} \sum_{\substack{\ell=1 \\ (\ell\neq k)}}^{M} W_{jk\ell} \Big] + \Big[S_{jk}^{2} \sum_{\substack{\ell=1 \\ (\ell\neq k)}}^{M} W^{*}_{jk\ell} \Big] \right\}$$

$$\text{Ak, A}\ell\text{:}\quad -2 \sum_{j=1}^{N} \Big[B_{jk} B_{j\ell} W_{jk\ell} + S_{jk} S_{j\ell} W^{*}_{jk\ell} \Big]$$

$$\text{Ak, Ck:}\quad 2 \sum_{j=1}^{N} \Big[B_{jk} \sum_{\substack{\ell=1 \\ (\ell\neq k)}}^{M} W_{jk\ell} \Big]$$

$$\text{Ak, C}\ell\text{:}\quad -2 \sum_{j=1}^{N} B_{j\ell} W_{jk\ell}$$

$$Ck,\ Ck:\quad 2 \sum_{j=1}^{N} \sum_{\substack{\ell=1 \\ (\ell \neq k)}}^{M} W_{jk\ell}$$

$$Ck,\ Cl:\quad -2 \sum_{j=1}^{N} W_{jk\ell}$$

AN UNWEIGHTED LEAST-SQUARES APPROXIMATION

An unweighted solution using information from item threshold estimates only may be obtained by redefining the weight terms in the fitting function F. First, all weights relating to item dispersion terms, $W^*jk\ell$, are set to zero. Second the weights relating to thresholds are replaced by simple indicator variables:

$$Djk\ell = \begin{cases} 1 & \text{if Item j is included in both calibrations,} \\ 0 & \text{otherwise.} \end{cases}$$

FINAL ESTIMATES OF ITEM PARAMETERS

The transformations determining the minimum of the fitting function will take group-effect estimates to the common scale defined by the first calibration. Item parameters may also be rescaled accordingly. Each item will have at least two estimates of its threshold and dispersion, accompanied by rescaled standard errors of estimation. (The standard errors of a rescaled item parameters is simply the standard error from the calibration run, multiplied by the appropriate transformation parameter Ak.) To obtain a single point estimate of a given parameter, one may take the average of the several estimates, each weighted by the squared reciprocal of its rescaled standard error.

With either the weighted or the unweighted solution, one may obtain an approximate Chi-square value to test the hypothesis that all estimates of a given item parameter are equivalent within the ranges of calibration error. For example, the Chi-square for the equality of the several estimates of the threshold of Item i is given by

$$\chi^2_{Bj} = \sum_{k=1}^{M} \delta_{jk} \left[\frac{(B_{jk}^{*} - B_{j.}^{*})^2}{\sqrt{(\sum_{\ell} \delta_{j\ell} \cdot SE^2(B_{j\ell}^{*}))/(\sum_{\ell} \delta_{j\ell} - 1)}} \right] ,$$

where $\delta jk=1$ if Item j was included in calibration k and 0 if not. The number of degrees of freedom for this quantity is the count of appearances of the item in all calibrations, minus one.

A test of fit for the entire set of linking transformations may be obtained by summing quantities as defined above, over all items and both thresholds and dispersions, with degrees of freedom similarly summed.

APPENDIX C

A COMPUTER PROGRAM FOR LINKING CALIBRATIONS

A FORTRAN IV COMPUTER PROGRAM FOR LINKING ITEM CALIBRATIONS

(SOURCE CODE AND EXAMPLE FROM 'UNDERSTANDING MATHEMATICAL CONCEPTS')

```
//CONCEPT JOB (8UZ303,NAEP,M),MISLEVY,RE=280K,TE=Y
//  EXEC FORTGCLG,USERLIB='SYS2.MATCAL'
//FORT.SYSIN DD *
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      NAMELIST/INPUT/ACRIT,N,M,METHOD,IDIAG,MAXITR,INAME
C
C                                      METHOD OF SOLUTION:
C                                      1. UNWEIGHTED LEAST SQUARES
C                                      2. WEIGHTED, THRESH INFO ONLY
C                                      3. WEIGHTED, THRESH & DISP INFO
C
C                                      N = TOTAL # ITEMS
C                                      M = # CALIBRATIONS
C                                      NP = # PARAMETERS TO BE ESTIMATED,
C                                           2*(M-1).
      METHOD=0
      MAXITR=10
      ACRIT=.001
      READ(5,INPUT)
C
      M1=M-1
      NP=2*M1
      NM=N*M
      NTRIS=(M1*(M1+1))/2
      NTRIL=(NP*(NP+1))/2
      CALL COMPUT
      STOP
      END
      SUBROUTINE COMPUT
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      DIMENSION INCID(20,4),B(20,4),BSE(20,4),S(20,4),SSE(20,4),
     $  RSCSLO(4),RSCINT(4),PARAMS(6),CHANGE(6),FDRV(6),
     $  SDRV(21),KDELTA(20,4,4),WB(20,4,4),WS(20,4,4),
     $  WORK1(6),WORK2(6),
     $  AVEB(20),AVES(20),ADJB(20),ADJS(20),AVEBSE(20),AVESSE(20),
     $  STNRES(80),SLOPE(20,4),SLOSE(20,4),AVESLO(20),AVSLSE(20)
      REAL*4 FMT(20)
C
C
      DO 10 K=1,M
        RSCSLO(K)=1D0
        RSCINT(K)=0D0
        DO 10 J=1,N
          INCID(J;K)=0
          B    (J,K)=0D0
          BSE  (J,K)=0D0
          S    (J,K)=0D0
          SSE  (J,K)=0D0
          SLOPE(J,K)=1E69
          SLOSE(J,K)=1E69
          DO 10 L=1,M
            KDELTA(J,K,L)=0
            WB    (J,K,L)=0D0
            WS    (J,K,L)=0D0
   10 CONTINUE
C
      READ(5,15)FMT
   15 FORMAT(20A4)
C
      DO 20 I=1,999999
        READ(5,FMT,END=21) J,K,THR,THRSE,DISP,DISPSE
        INCID(J,K)=1
        B  (J,K)=THR
        BSE(J,K)=THRSE
        S  (J,K)=DISP
        SSE(J,K)=DISPSE
   20 CONTINUE
   21 IF(IDIAG.LE.0) GOTO 30
      DO 30 K=1,M
        WRITE(6,9040) K
        DO 25 J=1,N
          IF(INCID(J,K).EQ.0) GOTO 22
          WRITE(6,9060) J,B(J,K),BSE(J,K),S(J,K),SSE(J,K)
   22     CONTINUE
   25   CONTINUE
   30 CONTINUE
C
C                                      SET UP KRONECKER DELTA MATRIX;
C                                      KDELTA(J,K,L)=1 IF ITEM J APPEARS
C                                      FOR BOTH CALIBRATIONS K & L,
C                                      =0 IF NOT.
   50 CONTINUE
      DO 60 J=1,N
        DO 60 K=1,M
          DO 60 L=1,M
            IF(INCID(J,K).EQ.1 .AND. INCID(J,L).EQ.1) KDELTA(J,K,L)=1
            WB(J,K,L)=KDELTA(J,K,L)
   60 CONTINUE
      IF(IDIAG.LT.2) GOTO 70
        DO 70 J=1,N
          WRITE(6,6000) J,(L,L=1,M)
          DO 65 K=1,M
            WRITE(6,6100)(K,(KDELTA(J,K,L),L=1,M))
```

```
  65      CONTINUE
6000     FORMAT('-KDELTA MATRIX, ITEM',I4,5X,20I3)
6100     FORMAT(22X,4X,21I3)
  70 CONTINUE
C
C
     ICYCL=-1
 100 ICYCL=ICYCL+1
     DO 700 ITR=1,MAXITR
       IF(IDIAG.GT.0) WRITE(6,9000) ICYCL,ITR
       DO 103 I=1,M1
         PARAMS(I)=RSCSLO(I+1)
         PARAMS(I+M1)=RSCINT(I+1)
 103   CONTINUE
C
C                                           IF WEIGHTED SOLUTION,
C                                           COMPUTE WEIGHTS.
C
       IF(METHOD.GE.1 .AND. ICYCL.GT.0)
    $    CALL WEIGHT(KDELTA,RSCSLO,BSE,WB)
       IF(METHOD.GE.2 .AND. ICYCL.GT.0)
    $    CALL WEIGHT(KDELTA,RSCSLO,SSE,WS)
C
C                                           CALCULATE DERIVATIVES
C
       CALL FIRST(B,BSE,S,SSE,RSCSLO,RSCINT,WB,WS,FDRV)
       CALL SECOND(B,BSE,S,SSE,RSCSLO,RSCINT,WB,WS,SDRV)
C
C                                           NEWTON-RAPHSON STEP
C
       CALL INVSD(SDRV,NP,DET,WORK1,WORK2)
       CALL MP'M(SDRV,FDRV,CHANGE,NP,NP,1,0,1)
       BIGC=0D0
       BIGD=0D0
       DO 150 I=1,NP
         CHANGE(I)=-CHANGE(I)
         IF(DABS(CHANGE(I)).GT.BIGC) BIGC=DABS(CHANGE(I))
         IF(DABS(FDRV(I)).GT.BIGD) BIGD=DABS(FDRV(I))
 150   CONTINUE
       CALL ADDM(PARAMS,CHANGE,PARAMS,NP,1,0)
       DO 160 I=1,M1
         RSCSLO(I+1)=PARAMS(I)
         RSCINT(I+1)=PARAMS(I+M1)
 160   CONTINUE
       CALL FUNCT(RSCSLO,RSCINT,B,S,WB,WS,CHISQ)
       WRITE(6,9020)ICYCL,ITR,BIGC,CHISQ
       IF(IDIAG.GT.0) CALL DPRNT(RSCSLO,1,M,0,8HSLOPES   )
       IF(IDIAG.GT.0) CALL DPRNT(RSCINT,1,M,0,8HINTERCPT)
       IF((BIGC.LE.ACRIT .OR. BIGD.LE.ACRIT) .OR.
    $    (METHOD.GT.2 .AND. ICYCL.EQ.0 .AND. ITR.GE.3)) GOTO 710
 700   CONTINUE
 710   CALL DPRNT(RSCSLO,1,M,0,8HSLOPES   )
       CALL DPRNT(RSCINT,1,M,0,8HINTERCPT)
```

```
      CALL DPRNT(SDRV,NP,NP,1,8HCOVARNCE)
C
C                                      IF WEIGHTED SOLUTION DESIRED,
C                                      AND JUST UNWEIGHTED SOLUTION
C                                      HAS BEEN COMPUTED, GO BACK
C                                      AND DO WEIGHTED SOLUTION.
C
      IF (ICYCL.EQ.0 .AND. METHOD.GE.1) GOTO 100
C
C                                      RESCALE ITEM PARAMETERS
C
      DO 765 J=1,N
C
        AVEB  (J)=0D0
        AVES  (J)=0D0
        ADJB  (J)=0D0
        ADJS  (J)=0D0
        AVEBSE(J)=0D0
        AVESSE(J)=0D0
  765 CONTINUE
      DO 780 K=1,M
        WRITE(6,9050) K
        DO 775 J=1,N
          IF(INCID(J,K).EQ.0) GOTO 770
C
            B  (J,K)= B  (J,K)*RSCSLO(K) + RSCINT(K)
            BSE(J,K)= BSE(J,K)*RSCSLO(K)
            S  (J,K)= S  (J,K)*RSCSLO(K)
            SSE(J,K)= SSE(J,K)*RSCSLO(K)
            SLOPE(J,K)=1D0/S(J,K)
            SLOSE(J,K)=SSE(J,K)*SLOPE(J,K)**2
            WGTB=0D0
            WGTS=0D0
            IF(BSE(J,K).GT.0D0) WGTB = 1D0/(BSE(J,K)**2)
            IF(SSE(J,K).GT.0D0) WGTS = 1D0/(SSE(J,K)**2)
            AVEB(J)=AVEB(J)+B(J,K)*WGTB
            AVES(J)=AVES(J)+S(J,K)*WGTS
            ADJB(J)=ADJB(J)+WGTB
            ADJS(J)=ADJS(J)+WGTS
            AVEBSE(J)=AVEBSE(J) + (WGTB*BSE(J,K))**2
            AVESSE(J)=AVESSE(J) + (WGTS*SSE(J,K))**2
          WRITE(6,9060) J,B(J,K),BSE(J,K),S(J,K),SSE(J,K),
     $      SLOPE(J,K),SLOSE(J,K)
  770     CONTINUE
  775   CONTINUE
  780 CONTINUE
      WRITE(6,9070)
      DO 830 J=1,N
        AVESLO(J)=0D0
        AVSLSE(J)=0D0
        IF(ADJB(J).GT.0D0) AVEB(J)=AVEB(J)/ADJB(J)
        IF(ADJS(J).GT.0D0) AVES(J)=AVES(J)/ADJS(J)
        IF(ADJB(J).GT.0D0) AVEBSE(J)=(1D0/ADJB(J))*DSQRT(AVEBSE(J))
        IF(ADJS(J).GT.0D0) AVESSE(J)=DSQRT(1D0/AVESSE(J))
        IF(AVES(J).NE.0D0) AVESLO(J)=1D0/AVES(J)
        IF(AVES(J).NE.0D0) AVSLSE(J)=AVESSE(J)*AVESLO(J)**2
        WRITE(6,9060) J,AVEB(J),AVEBSE(J),AVES(J),AVESSE(J),
     $    AVESLO(J),AVSLSE(J)
  830 CONTINUE
C
C                                                   STANDARDIZED RESIDUALS
C
      IDF=0
      DO 840 K=2,M
        K1=K-1
        DO 837 L=1,K1
          DO 834 J=1,N
            IDF=IDF+KDELTA(J,K,L)
  834     CONTINUE
  837   CONTINUE
  840 CONTINUE
C
      CHISQ=0D0
      DO 850 J=1,N
        DO 845 K=1,M
          INDX=J + (K-1)*N
          STNRES(INDX)=0D0
          IF(INCID(J,K).LE.0) GOTO 845
          STNRES(INDX)=(B(J,K)-AVEB(J))/DSQRT(AVEBSE(J)**2+BSE(J,K)**2)
          CHISQ=CHISQ + STNRES(INDX)**2
  845   CONTINUE
  850 CONTINUE
      CALL DPRNT(STNRES,N,M,0,8HRESIDUAL)
      WRITE(6,8000)
      DO 900 J=1,N
        WRITE(6,8100) INAME(J),
     $    (B(J,K),BSE(J,K),SLOPE(J,K),SLOSE(J,K),K=1,4),
     $     AVEB(J),AVEBSE(J),AVESLO(J),AVSLSE(J)
  900 CONTINUE
C
 8000 FORMAT(1H1//'   ITEM   ',5(' THRESH  SE  SLOPE  SE ')/
     $   1X,31(4H----))
 8100 FORMAT(1X,A8,5(F7.2,F5.2,F6.2,F5.2))
 9000 FORMAT(1H1//' CYCLE',I4,'     ITERATION',I4)
 9020 FORMAT(1H ,2I4,2F12.6)
 9040 FORMAT(1H1/'-INPUT ITEM PARAMETERS FOR CALIBRATION',I4//
     $  ' ITEM THRESHOLD   S.E.   DISPERSION   S.E.   '/
     $  ' --------------------------------------------')
 9050 FORMAT(1H1/'-RESCALED ITEM PARAMETERS FOR CALIBRATION',I4//
     $  ' ITEM THRESHOLD   S.E.   DISPERSION   S.E.   ',
     $        '   SLOPE    S.E.'/
     $  ' --------------------------------------------',
     $        '--------------------')
 9060 FORMAT(1X,I4,6F10.3)
 9070 FORMAT(1H1//'-GRAND AVERAGES OF ITEM PARAMETERS'/
     $  ' ITEM THRESHOLD   S.E.   DISPERSION   S.E.   ',
     $          '   SLOPE    S.E.'/
     $ ' -----------------------------------------------',
     $         '--------------------')
      RETURN
      END
      SUBROUTINE WEIGHT(KDELTA,RSCSLO,SE,W)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      DIMENSION KDELTA(20,4,4),RSCSLO(4),SE(20,4),W(20,4,4)
C
      DO 400 K=1,M
        K1=K-1
        DO 300 L=1,K1
          DO 200 J=1,N
            W(J,K,L)=0D0
            IF(KDELTA(J,K,L).LE.0) GOTO 100
            W(J,K,L)=1D0/((RSCSLO(K)*SE(J,K))**2+(RSCSLO(L)*SE(J,L))**2)
  100       W(J,L,K)=W(J,K,L)
  200     CONTINUE
  300   CONTINUE
  400 CONTINUE
      IF(IDIAG.LT.2) GOTO570
        DO 570 J=1,N
          WRITE(6,6000) J,(L,L=1,M)
          DO 565 K=1,M
            WRITE(6,6100)(K,(W(J,K,L),L=1,M))
  565     CONTINUE
 6000   FORMAT('-WEIGHT MATRIX, ITEM',I4,10I10)
 6100   FORMAT(12X,4X,I10,10F10.4)
  570 CONTINUE
C
      RETURN
      END
      SUBROUTINE FUNCT(RSCSLO,RSCINT,B,S,WB,WS,CHISQ)
C
C                                                COMPUTE CHI-SQUARE
C
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      DIMENSION RSCSLO(4),RSCINT(4),B(20,4),S(20,4),
     $  WB(20,4,4),WS(20,4,4)
C
      CHISQ=0D0
      DO 500 K=1,M
        K1=K-1
        DO 400 L=1,K1
          IF(K1.LE.0) GOTO 400
          DO 300 J=1,N
            IF(WB(J,K,L).LE.0D0) GOTO 300
            CHISQ=CHISQ + (RSCSLO(K)*B(J,K) + RSCINT(K)
     $                     -RSCSLO(L)*B(J,L) - RSCINT(L))**2 * WB(J,K,L)
            IF(METHOD.LT.2) GOTO 300
            CHISQ=CHISQ + (RSCSLO(K)*S(J,K)
     $                     -RSCSLO(L)*S(J,L))**2 * WS(J,K,L)
  300     CONTINUE
  400   CONTINUE
  500 CONTINUE
      RETURN
      END
      SUBROUTINE FIRST(B,BSE,S,SSE,RSCSLO,RSCINT,WB,WS,FDRV)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      DIMENSION B(20,4),S(20,4),BSE(20,4),SSE(20,4),RSCSLO(4),RSCINT(4),
     $  WB(20,4,4),WS(20,4,4),FDRV(6)
C
C                                             FIRST DERIVS OF SLOPES
C
      DO 190 K=2,M
        KK=K-1
        FDRV(KK)=0D0
        DO 170 J=1,N
          DO 150 L=1,M
            IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 150
            FDRV(KK)=FDRV(KK) + WB(J,K,L)*
     $        (RSCSLO(K)*B(J,K)**2 - RSCSLO(L)*B(J,K)*B(J,L)
     $         + RSCINT(K)*B(J,K) - RSCINT(L)*B(J,K))
            IF(METHOD.LT.2) GOTO 150
            FDRV(KK)=FDRV(KK) + WS(J,K,L)*
     $        (RSCSLO(K)*S(J,K)**2 - RSCSLO(L)*S(J,K)*S(J,L))
  150     CONTINUE
  170   CONTINUE
        FDRV(KK)= FDRV(KK) * 2D0
  190 CONTINUE
C
C                                         FIRST DERIVS OF INTRCPS
C
      DO 290 K=2,M
        KK=M1 + (K-1)
        FDRV(KK)=0D0
        DO 270 J=1,N
          DO 250 L=1,M
            IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 250
            FDRV(KK)=FDRV(KK) +  WB(J,K,L)*
     $       (RSCSLO(K)*B(J,K) - RSCSLO(L)*B(J,L)
     $       + RSCINT(K) - RSCINT(L))
  250     CONTINUE
  270   CONTINUE
        FDRV(KK)= FDRV(KK) * 2D0
  290 CONTINUE
      IF(IDIAG.GT.0) CALL DPRNT(FDRV,1,NP,0,8HFDRV    )
```

```
C
      RETURN
      END
      SUBROUTINE SECOND(B,BSE,S,SSE,RSCSLO,RSCINT,WB,WS,SDRV)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 INAME(20)
      COMMON/PARCOM/INAME,ACRIT,N,M,NM,METHOD,M1,NP,NTRIS,NTRIL,
     $               IDIAG,MAXITR
      DIMENSION B(20,4),S(20,4),BSE(20,4),SSE(20,4),RSCSLO(4),RSCINT(4),
     $ WB(20,4,4),WS(20,4,4),SDRV(21),SDRVA(6),
     $ SDRVB(9),SDRVC(6)
C
C                                           SLOPE DOUBLE DERIVS
C
      INDX=0
      DO 190 K=2,M
        INDX=INDX + (K-1)
        SDRVA(INDX)=0D0
        DO 170 J=1,N
          SUMWB=0D0
          SUMWS=0D0
          DO 150 L=1,M
            IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 150
            SUMWB=SUMWB + WB(J,K,L)
            SUMWS= SUMWS + WS(J,K,L)
  150     CONTINUE
          SDRVA(INDX)=SDRVA(INDX)+ SUMWB*B(J,K)**2
          IF(METHOD.GE.2)
     $      SDRVA(INDX)=SDRVA(INDX) + SUMWS*S(J,K)**2
  170   CONTINUE
        SDRVA(INDX)=SDRVA(INDX) * 2D0
  190 CONTINUE
C
C                                          SLOPE CROSS DERIVS
C
      INDX=0
      DO 290 K=2,M
        DO 270 L=2,K
          INDX=INDX+1
          IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 270
          SDRVA(INDX) = 0D0
          DO 250 J=1,N
            SDRVA(INDX)=SDRVA(INDX) + B(J,K)*B(J,L)*WB(J,K,L)
            IF(METHOD.LT.2) GOTO 250
            SDRVA(INDX)=SDRVA(INDX) + S(J,K)*S(J,L)*WS(J,K,L)
  250     CONTINUE
          SDRVA(INDX) = - SDRVA(INDX) * 2D0
  270   CONTINUE
  290 CONTINUE
C
C                                             SLOPE*INTRCP CROSS DERIVS,
C                                             SAME CALIBRATION
C
```

```
      DO 390 K=2,M
        KK=K-1
        INDX=(KK-1)*M1 + KK
        SDRVB(INDX)=0D0
        DO 370 J=1,N
          SUMWB = 0D0
          DO 350 L=1,M
            IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 350
            SUMWB=SUMWB + WB(J,K,L)
  350     CONTINUE
          SDRVB(INDX)=SDRVB(INDX) + B(J,K)*SUMWB
  370   CONTINUE
        SDRVB(INDX) = SDRVB(INDX) * 2D0
  390 CONTINUE
C
C                                            SLOPE*INTRCP CROSS DERIVS,
C                                            DIFFERENT CALIBRATIONS
C
      INDX=0
      DO 490 K=2,M
        DO 470 L=2,M
          INDX=INDX+1
          IF(K.EQ.L) GOTO 470
          SDRVB(INDX)=0D0
          DO 450 J=1,N
            SDRVB(INDX)=SDRVB(INDX) + B(J,K)*WB(J,K,L)
  450     CONTINUE
          SDRVB(INDX) = - SDRVB(INDX) * 2D0
  470   CONTINUE
  490 CONTINUE
C
C                                            INTRCP DOUBLE DERIVS
C
      INDX=0
      DO 590 K=2,M
        INDX=INDX + (K-1)
        SDRVC(INDX) = 0D0
        DO 570 J=1,N
          DO 550 L=1,M
            IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 550
            SDRVC(INDX) = SDRVC(INDX) + WB(J,K,L)
  550     CONTINUE
  570   CONTINUE
        SDRVC(INDX) = SDRVC(INDX) * 2D0
  590 CONTINUE
C
C                                            INTRCP CROSS DERIVS
C
      INDX=0
      DO 690 K=2,M
        DO 670 L=2,K
          INDX=INDX + 1
          IF(L.EQ.K .OR. WB(J,K,L).LE.0D0) GOTO 670
```

```
     SDRVC(INDX) = 0D0
     DO 650 J=1,N
      SDRVC(INDX) = SDRVC(INDX) + WB(J,K,L)
 650 CONTINUE
     SDRVC(INDX) = - SDRVC(INDX) * 2D0
 670 CONTINUE
 690 CONTINUE
C
   CALL ADJRC(SDRVA,SDRVB,SDRVC,SDRV,M1,M1)
   IF(IDIAG.GT.0) CALL DPRNT(SDRV,NP,NP,1,8HSDRV    )
   RETURN
   END
//GO.SYSIN DD *
 &INPUT N=17,M=4,METHOD=3,MAXITR=20,INAME=
 '5-A45532',
 '5-B41532',
 '5-B41732',
 '5-B31732',
 '5-N00002',
 '5-B11008',
 '5-A71043',
 '5-A21022',
 '5-B32632',
 '5-K30004',
 '5-K10010',
 '5-B33232',
 '5-G43009',
 '5-H12025',
 '5-G20001',
 '5-K51020',
 '5-A21032',
 &END
(I2,I1,4F8.3)
 41   3.046    .360   5.263    .281
 51  -1.687    .360   5.000    .750
 61    .185    .220   6.250    .781
 81    .413    .160   4.000    .480
 91   2.681    .320   5.556    .926
101   1.107    .270  10.000   2.000
111   3.756    .440   5.263    .831
121    .227    .220   6.667    .889
131   1.923    .280   7.143   1.020
 52  -3.460    .610   5.000    .750
 62  -3.010    .720  10.000   2.000
102    .810    .290   7.692   1.183
112   3.590    .610   5.556    .926
142    .700    .290   7.692   1.183
162   -.150    .210   4.762    .680
 13  -1.511    .205   2.857    .335
 23   -.413    .210   6.061    .882
 43   -.490    .195   5.348    .744
 53  -3.420    .478   4.115    .559
 63  -2.540    .360   4.184    .543
```

```
524.     73  -1.378    .231   4.065    .512
525.     83  -1.319    .194   3.040    .351
526.    103  -2.325    .357   4.717    .623
527.     64  -5.283    .968   6.944   1.206
528.     84  -2.657    .454   5.155    .797
529.    114    .791    .260   6.329   1.001
530.    144  -2.084    .397   6.289    .989
531.    154   2.233    .338   5.464    .836
532.    174   1.533    .324   6.667   1.067
```

APPENDIX D

COMMENTS ON NAEP PUBLIC-USE DATA TAPES

COMMENTS ON NAEP PUBLIC-USE DATA TAPES

Due to the quantity of information provided, the NAEP tapes were in general cumbersome to work with. The documentation for the 1977/1978 and change item tapes can only be described as excellent, comprehensive in scope and accurate in detail. The files comprising the tapes were well-organized, the information about variable locations as well as that contained in the value labels of the accompanying SPSS files was invaluable, and the classification of items contained in the appendices greatly facilitated the construction of our scales. In comparison, the 1972/1973 tape was more difficult to work with, the organization and contents of the tape less readily understood.

A few minor difficulties that we encountered bear mentioning. The difference between the "no response" and "missing values" classifications of item responses is unclear from the documentation. The fact that in school and out of school 17 year olds are assigned different values for the region variable proved to be a source of temporary confusion.

Data for items which were supposed to be invariant across two or more age/year combinations, according to the documentation provided, occasionally did not seem right. As an example, Item 5-B32632 appeared in both the 13-year old and 17-year old instruments in 1977/78, as T1020 and S0921 respectively. Proportions of correct response, however, suggested the item to be exteremly more difficult for 17-year olds than 13-year olds, a trend strongly contradicting evidence from every other item linking these two age years. It is likely that data for the 17-year olds is in error here. Similar problems arose for items 5-A71043 and 5-N00002. Such questionable item/age/year data combinations were omitted in our computations.